SU 1744

AGH

Ireneusz Czajka
Andrzej Gołaś

INŻYNIERSKIE METODY ANALIZY NUMERYCZNEJ I PLANOWANIE EKSPERYMENTU

SKRYPTY UCZELNIANE

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDAWNICTWA AGH

KRAKÓW 2017

SU 1744 pozycja wydawnictw dydaktycznych Akademii Górniczo-Hutniczej
im. Stanisława Staszica w Krakowie

© Wydawnictwa AGH, Kraków 2017

ISSN 0239-6114
ISBN 978-83-7464-941-4

Redaktor Naczelny Wydawnictw AGH: *Jan Sas*

Komitet Naukowy Wydawnictw AGH:
Andrzej Pach (przewodniczący)
Jan Chłopek
Barbara Gąciarz
Bogdan Sapiński
Stanisław Stryczek
Tadeusz Telejko

Recenzenci:
dr hab. inż. Adam Brański, prof. nadzw. Politechniki Rzeszowskiej
dr hab. inż. Marek Iwaniec, prof. nadzw. Akademii Górniczo-Hutniczej

Afiliacja autorów:
AGH Akademia Górniczo-Hutnicza

Opieka redakcyjna:
Małgorzata Koch

Skład komputerowy: *Ireneusz Czajka*

Redakcja Wydawnictw AGH
al. A. Mickiewicza 30, 30-059 Kraków
tel. 12 617 32 28, 12 636 40 38
e-mail: redakcja@wydawnictwoagh.pl
http://www.wydawnictwo.agh.edu.pl

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Spis treści

Od autorów 11

I. Klasyczne metody numeryczne

Rozdział 1. Na początek 15
1.1. Wstęp z odrobiną historii 15
1.2. Czym są metody numeryczne 17
1.3. Błędy metod numerycznych 18
1.4. Poruszane problemy 24

Rozdział 2. Interpolacja 25
2.1. Interpolacja wielomianowa 26
2.1.1. Interpolacja wielomianami Czebyszewa 27
2.1.2. Przykład obliczeniowy 29
2.1.3. Interpolacja Newtona 29
2.1.4. Przykład obliczeniowy 31
2.1.5. Przykład obliczeniowy 32
2.1.6. Interpolacja Lagrange'a 34
2.1.7. Przykład obliczeniowy 35
2.2. Podsumowanie 36

Rozdział 3. Rozwiązywanie równań nieliniowych 37
3.1. Wstęp 37
3.2. Metoda iteracji prostej 37
3.2.1. Interpretacja geometryczna 39
3.2.2. Przykład obliczeniowy 39
3.3. Nadrelaksacja – poprawianie zbieżności 43
3.3.1. Przykład obliczeniowy 44
3.3.2. Przykład obliczeniowy 45
3.4. Metoda Δ^2 Aitkena 46
3.5. Związek metody iteracji prostej z metodą Newtona 47
3.6. Metoda Newtona 48
3.6.1. Interpretacja geometryczna 48
3.6.2. Przykład obliczeniowy 50
3.6.3. Zbieżność metody 50
3.7. Metoda siecznych 52
3.7.1. Przykład obliczeniowy 53
3.8. Regula falsi 54
3.9. Warunki zakończenia obliczeń 56
3.10. Przykłady praktyczne 56
3.10.1. Pomiar strumienia objętości za pomocą kryzy 57

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

3.10.2. Tok obliczeń . . . 57
3.10.3. Rozwiązanie . . . 60
3.10.4. Przykład obliczeniowy . . . 62
3.11. Miejsca zerowe wielomianów . . . 64
3.11.1. Metoda iteracyjna Laguerre'a . . . 65
3.12. Rozwiązywanie równań nieliniowych z wykorzystaniem MATLAB-a . . . 65
3.13. Podsumowanie . . . 66

Rozdział 4. Całkowanie numeryczne . . . 67

4.1. Wstęp . . . 67
4.2. Metoda prostokątów . . . 68
4.3. Metoda trapezów . . . 70
4.3.1. Ekstrapolacja Richardsona w metodzie trapezów . . . 71
4.4. Metoda Romberga . . . 73
4.4.1. Przykład obliczeniowy . . . 74
4.5. Wzór Simpsona . . . 75
4.5.1. Przykład obliczeniowy . . . 78
4.6. Kwadratury Newtona–Cotesa . . . 80
4.7. Kwadratury Gaussa . . . 82
4.7.1. Kwadratury Gaussa–Kronroda . . . 86
4.7.2. Kilka słów podsumowania . . . 88
4.8. Metody Monte Carlo . . . 88
4.8.1. Przykład obliczeniowy . . . 89
4.9. Przykłady praktyczne . . . 91
4.9.1. Moc sprężarki . . . 91
4.9.2. Wymiana ciepła . . . 97
4.10. Całkowanie w pakietach MATLAB i GNU Octave . . . 98
4.11. Podsumowanie . . . 99

Rozdział 5. Różniczkowanie numeryczne . . . 100

5.1. Wyznaczanie wartości pochodnej . . . 101
5.1.1. Różniczkowanie trójpunktowe . . . 103
5.1.2. Przykład obliczeniowy . . . 108
5.2. Inne metody różniczkowania . . . 110
5.2.1. Różniczkowanie sygnałów zaszumionych . . . 110
5.3. Podsumowanie . . . 111

Rozdział 6. Rozwiązywanie równań różniczkowych . . . 112

6.1. Podstawowe informacje o równaniach różniczkowych . . . 112
6.2. Metody rozwiązywania problemów początkowych . . . 117
6.2.1. Równania różniczkowe wyższych rzędów . . . 118
6.2.2. Metody jednokrokowe . . . 120
6.2.3. Metoda Rungego–Kutty rzędu drugiego . . . 128
6.2.4. Przykład obliczeniowy . . . 129
6.2.5. Przykład . . . 133
6.2.6. Metoda Rungego-Kutty rzędu czwartego . . . 134
6.2.7. Ogólne informacje o metodach Rungego–Kutty . . . 136
6.2.8. Tablice Butchera . . . 137
6.2.9. Metody Rungego–Kutty z estymacją błędu . . . 138
6.2.10. Niejawne metody Rungego–Kutty . . . 139

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.2.11. Metody wielokrokowe . . . 140
6.2.12. Niejawna metoda trapezów . . . 141
6.3. Szacowanie lokalnego błędu rozwiązania . . . 147
6.4. Metody rozwiązywania problemów brzegowych . . . 148
6.5. Wykorzystanie funkcji MATLAB-a . . . 149
6.5.1. Przykład obliczeniowy . . . 149
6.6. Podsumowanie . . . 150

Rozdział 7. Rozwiązywanie układów liniowych równań algebraicznych . . . 151

7.1. Wstęp . . . 151
7.1.1. Układy równań liniowych . . . 151
7.1.2. Normy wektorów i macierzy . . . 152
7.2. Metody bezpośrednie . . . 156
7.2.1. Eliminacja Gaussa . . . 156
7.2.2. Metoda Gaussa–Jordana . . . 159
7.3. Metody polegające na rozkładach macierzy . . . 159
7.3.1. Rozkład LU . . . 159
7.3.2. Rozkład QR . . . 161
7.4. Metody iteracyjne . . . 162
7.4.1. Metoda Gaussa–Seidla . . . 162
7.5. Metody najszybszego spadku i gradientu sprzężonego . . . 165
7.5.1. Metoda najszybszego spadku . . . 165
7.5.2. Metoda gradientu sprzężonego . . . 166
7.5.3. Przyspieszanie obliczeń . . . 168
7.6. Podsumowanie . . . 169

Rozdział 8. Aproksymacja . . . 170

8.1. Aproksymacja punktowa wielomianowa . . . 171
8.1.1. Aproksymacja średniokwadratowa . . . 171
8.1.2. Przykład praktyczny . . . 175
8.1.3. Rozwiązanie . . . 176
8.2. Podsumowanie . . . 179

II. Zaawansowane metody numeryczne

Rozdział 9. Metoda różnic skończonych . . . 183

9.1. Wstęp . . . 183
9.2. Pochodna pierwszego rzędu . . . 184
9.3. Pochodne wyższych rzędów . . . 185
9.4. Pochodne mieszane . . . 186
9.5. Przykład dwuwymiarowej analizy . . . 188
9.6. Błędy . . . 192

Rozdział 10. Metoda elementów brzegowych . . . 193

10.1. Wstęp . . . 193
10.1.1. Metoda Trefftza . . . 194
10.2. Podstawy matematyczne . . . 195
10.3. Równania całkowe metody elementów brzegowych . . . 196
10.3.1. Ustalony przepływ ciepła . . . 196
10.3.2. Numeryczna aproksymacja równań brzegowych . . . 199
10.4. Przykład obliczeniowy . . . 203

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

10.5. Podsumowanie . . . 210
Rozdział 11. Metoda elementów skończonych . . . 211
11.1. Wstęp . . . 211
11.2. Idea metody elementów skończonych . . . 211
11.3. Matematyczny opis układów mechanicznych . . . 214
11.4. Przybliżone metody rozwiązywania równań . . . 214
11.4.1. Metoda Rayleigha–Ritza . . . 214
11.4.2. Metoda Galerkina . . . 219
11.5. Wyznaczenie macierzy sztywności elementu prętowego . . . 225
11.6. Wyznaczenie macierzy bezwładności elementu prętowego . . . 229
11.7. Agregacja macierzy sztywności i bezwładności . . . 230
11.7.1. Nakładanie warunków brzegowych Dirichleta . . . 231
11.7.2. Przykład rozwiązania jednowymiarowego problemu statycznego . . . 233
11.7.3. Macierz sztywności i bezwładności w globalnym układzie odniesienia . . . 236
11.8. Elementy wyższych rzędów . . . 241
11.9. Rozwiązywanie prostych zagadnień statyki . . . 243
11.9.1. Przykład statyczny płaski . . . 243
11.10. Zagadnienia dynamiczne . . . 249
11.10.1. Superpozycja modalna . . . 252
11.10.2. Przykład analizy drgań swobodnych pręta . . . 255
11.10.3. Przykład analizy dynamicznej . . . 259
11.11. Macierz sztywności i bezwładności elementu belkowego . . . 263
11.12. Podsumowanie i kilka słów o błędach . . . 268
Rozdział 12. Obliczenia symboliczne . . . 270
12.1. Wstęp . . . 270
12.2. Maxima . . . 270
12.2.1. Wyrażenia i funkcje . . . 271
12.2.2. Różniczkowanie i całkowanie . . . 272
12.2.3. Macierze . . . 273
12.2.4. Równania i układy równań . . . 274
12.2.5. Równania różniczkowe . . . 275
12.2.6. Pozostałe możliwości . . . 276
12.2.7. Wybrane funkcje . . . 279
12.3. SageMath . . . 283
12.3.1. Wyrażenia i funkcje . . . 283
12.3.2. Różniczkowanie i całkowanie . . . 284
12.3.3. Macierze . . . 285
12.3.4. Równania i układy równań . . . 286
12.3.5. Równania różniczkowe . . . 288
12.4. Podsumowanie . . . 288

III. Metody statystyczne

Rozdział 13. Metody statystyczne . . . 293
13.1. Wstęp . . . 293
13.2. Pojęcia podstawowe . . . 294
13.2.1. Prawdopodobieństwo . . . 294
13.2.2. Rozkłady gęstości prawdopodobieństwa zmiennych losowych . . . 298

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

13.2.3. Obiekt i jego model . . . 301
13.2.4. Optymalizacja . . . 302
13.2.5. Eksperyment i jego plan . . . 303
13.3. Proste modele matematyczne obiektów . . . 306
13.4. Wybór zmiennych wejściowych . . . 307
13.5. Analiza regresji . . . 308
13.5.1. Wyznaczanie parametrów funkcji regresji . . . 311
13.5.2. Jakość funkcji regresji . . . 313
13.5.3. Przedziały ufności dla funkcji regresji . . . 316
13.5.4. Adekwatność funkcji regresji . . . 317
13.6. Podsumowanie . . . 319

Rozdział 14. Planowanie eksperymentu . . . 320

14.1. Wstęp . . . 320
14.2. Plany dwupoziomowe . . . 322
14.3. Plany dwupoziomowe ułamkowe . . . 328
14.4. Plany trójpoziomowe . . . 334
14.5. Plany Placketta–Burmana . . . 337
14.6. Plany wielopoziomowe . . . 337
14.6.1. Plany kompozycyjne . . . 338
14.6.2. Planowanie rotatabilne . . . 343
14.6.3. Plany Boxa–Behnkena . . . 345
14.7. Planowanie losowe . . . 346
14.8. Wspomaganie realizacji eksperymentów komputerowych . . . 346
14.9. Wykorzystanie pakietu DAKOTA . . . 347
14.10. Podsumowanie . . . 349

Bibliografia . . . 351

Skorowidz . . . 355

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Alicji,
która jest moim natchnieniem
Ireneusz Czajka

Żonie, dzieciom i wnukom
Andrzej Gołaś

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Serce rozumne zdobywa wiedzę,
a ucho mądrych dąży do wiedzy.
Księga Przysłów 18,15

Od autorów

Problemy inżynierskie mechaniki stosowanej dzielone są na trzy działy: projektowanie, wytwarzanie oraz eksploatacja. W każdym z nich analizuje się związki przyczynowo-skutkowe pomiędzy wielkościami fizycznymi opisującymi badany proces fizyczny. W zadaniach prostych, gdzie znamy przyczyny (wymuszenia), interesują nas nieznane skutki (odpowiedzi układu). W zadaniach odwrotnych poszukujemy takich wymuszeń, które doprowadzają do oczekiwanych odpowiedzi. Zarówno w zadaniach prostych, jak i odwrotnych, kluczowym problemem jest określenie relacji ilościowych pomiędzy zbiorami wielkości wejściowych i wyjściowych. Te relacje nazywamy modelem matematycznym. W zagadnieniach statyki i wytrzymałości materiałów są to na ogół układy algebraicznych równań liniowych i nieliniowych. W zagadnieniach dynamicznych mamy do czynienia z układami równań różniczkowych zwyczajnych (dla układów o parametrach skupionych) i równań różniczkowych cząstkowych dla układów o parametrach rozłożonych. Skuteczne rozwiązanie wyżej wymienionych układów równań wymaga spełnienia dwóch warunków:

— znajomości metod numerycznych,
— umiejętności posługiwania się współczesnymi narzędziami informatycznymi, obejmującymi sprzęt i oprogramowanie, umożliwiającymi zastosowanie metod numerycznych.

Metody numeryczne to dział matematyki opisujący sposoby rozwiązywania pewnych problemów. Z inżynierskiego punktu widzenia istotne są numeryczne metody aproksymacji, rozwiązywania algebraicznych równań liniowych i nieliniowych, równań różniczkowych zwyczajnych i cząstkowych. Przy badaniach identyfikacyjnych ważne są metody planowania eksperymentu i statystycznej analizy wyników. Chociaż wiele osób metody numeryczne jednoznacznie kojarzy z komputerami, to trzeba pamiętać, że metody numeryczne pojawiły się znacznie wcześniej niż współczesne komputery. Pierwsze takie metody, o których w tej chwili wiemy, tworzyli Babilończycy. Istniejące doniesienia wskazują na to, że matematycy chińscy posiadali sporą wiedzę w tej dziedzinie. Według dostępnych informacji prace hinduskie znacznie wyprzedzały osiągnięcia naukowców europejskich. Na podstawie dostępnych dowodów można stwierdzić, że wprowadzenie symbolu zera, które stanowiło przełom pozwalający na wprowadzenie pozycyjnego zapisu liczb, miało miejsce na terenach dzisiejszej Kambodży.

W erze nowożytnej metody numeryczne tworzyli m.in.: Isaac Newton, Joseph Louis Lagrange czy Carl Friedrich Gauss. Tworzyli procedury (dzisiaj nazywane algorytmami), których ówczesne wykorzystanie na pełną skalę było niemożliwe ze względu na brak narzędzi obliczeniowych. Dopiero stworzenie komputerów o odpowiedniej mocy

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

obliczeniowej daje możliwość wykorzystania geniuszu uczonych sprzed stuleci. Ten fakt miał duży wpływ na zawartość niniejszej książki.

Autorzy odstąpili od wyprowadzania wzorów, nie podają dowodów twierdzeń. Kładą natomiast nacisk na praktyczną ilustrację umiejętności zastosowania podanych algorytmów. Po każdym wprowadzeniu teoretycznym prezentowany jest przykład z wykorzystaniem pakietów informatycznych, takich jak MATLAB, SageMath, Maxima czy DAKOTA. Oprócz klasycznych metod numerycznych autorzy dają też przykłady zastosowań metod elementów skończonych i brzegowych, które należą do zaawansowanych metod numerycznych, ale też obecnie stają się metodami klasycznymi.

Książka adresowana jest przede wszystkim do studentów wydziałów mechanicznych, ale ze względu na uniwersalność zawartych w niej przykładów może być pomocna dla studentów szeroko pojętych studiów inżynierskich.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Część I

Klasyczne metody numeryczne

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 1

Na początek

1.1. Wstęp z odrobiną historii

Ze względu na ciągłą potrzebę podnoszenia jakości życia i wrodzoną ciekawość, ludzie od zawsze przejawiali chęć do tworzenia nowych konstrukcji. Wraz z podejmowanymi próbami konstruowania nowych obiektów technicznych pojawiały się problemy natury matematycznej wymagające rozwiązania. I tak można powiedzieć, że przy konstruowaniu ścian pojawił się problem wyznaczania długości przekątnej prostokąta czy kwadratu, a próba stworzenia mostu czy stropu wymagała wyznaczenia długości łuku. Jeszcze częściej budowniczowie trafiali na problem wyznaczenia kąta prostego. Chęć posiadania i świadomość własności sprawiły, że zaistniała potrzeba wyznaczania pola powierzchni posiadanego obszaru. Pojawiające się w ten sposób problemy rozwiązywano, wykorzystując istniejące ówcześnie możliwości techniczne i wiedzę matematyczną. Tak więc już w starożytnym Babilonie czy Chinach opracowywano „publikacje" opisujące sposoby rozwiązywania problemów technicznych. Na podstawie tych pierwszych „algorytmów" opracowywano narzędzia umożliwiające realizację obliczeń. W ten sposób, w starożytnym Babilonie powstały na przykład tablice iloczynowe czy tablice odwrotności[1], których funkcja sprowadzała się do ułatwienia złożonych obliczeń.

Najstarsze znane obecnie doniesienia dotyczące metod numerycznych wiążą się z bardzo wczesnym etapem rozwoju znanej cywilizacji. Jednym z przykładów obliczeń przybliżonych jest opisana na babilońskiej tabliczce glinianej, oznaczonej jako YBC 7289, metoda wyznaczania długości przekątnej kwadratu o długości boku równej jedności. Ten ważny z praktycznego punktu widzenia wynik został przedstawiony w systemie sześćdziesiątkowym jako

$$\sqrt{2} \approx 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} \approx 1,41421296$$

Błąd bezwzględny powyższego przybliżenia wynosi około $6 \cdot 10^{-7}$. Nawet współcześnie taki błąd jest akceptowalny w większości obliczeń inżynierskich, a samo oszacowanie pochodzi z lat 1800–1600 p.n.e.

Żyjący, prawdopodobnie, w pierwszym wieku n.e. w Aleksandrii grecki uczony Heron opisał iteracyjną metodę wyznaczania pierwiastka kwadratowego z liczby A. Po przyjęciu wartości początkowej x_0, metodę można zapisać jako

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{A}{x_i}\right)$$

[1] Tablice iloczynowe to na przykład VAT7858, NBC7344, MS2184/3 czy MS3866, zaś tablica odwrotności to na przykład MS3874.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Opisana formuła jest też nazywana metodą babilońską, ponieważ Heron nie wynalazł jej, a jedynie dokładnie opisał. Jak zostanie wykazane w dalszej części podręcznika, formuła Herona stanowi szczególny przypadek metody Newtona rozwiązywania równań nieliniowych.

Między II wiekiem p.n.e. a II wiekiem n.e, za czasów dynastii Han, w Chinach powstało dzieło *Dziewięć rozdziałów o sztukach matematycznych*, w którym omówiono metodę znaną obecnie jako schemat Hornera. Dopiero dużo później, bo w 1819 roku, William George Horner opisał metodę wyznaczania wartości wielomianu za pomocą znacznie mniejszej liczby działań (brak potęgowania) niż tradycyjnie. Chociaż metoda ta nosi nazwę schematu Hornera, w 1669 roku posługiwał się nią Isaac Newton. Jeszcze wcześniej, bo w XIII wieku, opisał ją chiński matematyk Qin Jiushao.

Prawdopodobnie w starożytnym Egipcie, a na pewno w starożytnej Grecji, znano trygonometrię. Najbardziej znanym i jednocześnie najdłużej używanym przyrządem wspomagającym obliczenia był abacus, którego daleki potomek nazywany jest dziś liczydłem.

Prawdziwy przełom nastąpił w roku 1614, kiedy to John Napier, jako pierwszy przedstawiciel zachodniej nauki, powtórzył odkrycie indyjskich uczonych z XIV wieku i sformułował pojęcie logarytmu. Już trzy lata później wynalazł przyrząd wspomagający obliczenia znany pod nazwą kostki Napiera.

Kolejnym przełomowym wynalazkiem był przyrząd pomysłu angielskiego matematyka Williama Oughtreda, który wykorzystując linijki logarytmiczne Edmunda Guntera, zbudował w 1622 roku instrument pozwalający na bezpośrednie mnożenie i dzielenie liczb, będący prekursorem znanego suwaka logarytmicznego. Niecałe 230 lat później francuski artylerzysta Amedee Mannheim opracował suwak logarytmiczny w znanym dziś kształcie. Ów pomysł z lat 50. XIX wieku na ponad 100 lat zmonopolizował obliczenia.

Jeszcze przed skonstruowaniem suwaka logarytmicznego podejmowano próby budowy mechanicznych urządzeń wspomagających obliczenia. W roku 1623 niemiecki astronom i matematyk Wilhelm Schickard na zamówienie Johannesa Keplera wykonał pierwsze mechaniczne urządzenie liczące. Urządzenie to, zwane zegarem liczącym, wykorzystywało pomysł Napiera, jednak Schickard zamiast kostek użył walców, co dawało możliwość wykonywania obliczeń cyklicznie. Maszyna mogła realizować cztery podstawowe działania: dodawanie, odejmowanie, mnożenie i dzielenie na liczbach całkowitych. Z kolei 22-letni Blaise Pascal około roku 1645 zbudował mechaniczny sumator, nazywany dziś pascaliną. Następnie w 1671 roku Gottfried Wilhelm Leibniz rozwinął obie konstrukcje w maszynę zwaną *lebendige Rechenbank*. Konstrukcja Leibniza oprócz dodawania, odejmowania, mnożenia oraz dzielenia, umożliwiała również wyliczanie wartości pierwiastków kwadratowych.

Naprawdę przełomową konstrukcją była maszyna różnicowa skonstruowana przez angielskiego matematyka i wynalazcę Charlesa Babbage'a w roku 1822. Celem Babbage'a było zbudowanie maszyny pozwalającej przyspieszyć tworzenie tablic matematycznych i astronomicznych. Chociaż Babbage nigdy nie skończył swojego projektu, jednak na jego podstawie w latach 1989–1991 w londyńskim Muzeum Nauki zbudowano działający egzemplarz maszyny różnicowej. W trakcie prac stosowano dokładność dostępną w XIX wieku, to pozwoliło udowodnić możliwość realizacji pomysłu

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Babbage'a za jego czasów. Urządzenie wykonywało obliczenia, wykorzystując liczby o długości 31 cyfr. W trakcie prac nad maszyną różnicową, Babbage wpadł na pomysł zbudowania znacznie bardziej zaawansowanej maszyny, nazwanej przez niego maszyną analityczną. Pierwszy projekt powstał w 1837 roku. Maszyna analityczna stanowiła pierwszy programowalny komputer ogólnego stosowania. Pierwszy program na tę właśnie maszynę (choć nigdy nie została zbudowana) w roku 1843 napisała Ada Lovelace.

W latach 40. XX wieku niemiecki inżynier Konrad Zuse zbudował serię elektromechanicznych maszyn liczących. Najbardziej znana z nich to Z3, która powstała w 1941 roku i jest uznawana za pierwszą w historii działającą programowalną maszynę liczącą.

W 1944 roku Howard Aiken na podstawie koncepcji maszyny różnicowej Babbage'a zbudował maszynę Harvard Mark I, która ważyła 5 ton, zawierała trzy miliony połączeń i 800 kilometrów przewodów elektrycznych. Mark I potrafił tablicować funkcje, całkować numerycznie, rozwiązywać równania różniczkowe i układy równań algebraicznych liniowych.

Wyprodukowany w roku 1971 4-bitowy procesor 4004 był pierwszym cywilnym jednoukładowym procesorem komputerowym. Dzięki tej konstrukcji nastąpił kolejny przełom w historii obliczeń numerycznych.

Oprócz działań na gruncie konstrukcyjnym, wiele wysiłku poświęcono metodom rozwiązywania pojawiających się problemów. Jak wcześniej wspomniano, w Babilonie i Chinach rozwiązywano praktyczne problemy. W Babilonie znano między innymi wartość $\sqrt{2}$ wyrażoną w postaci ułamka zwykłego, a za pomocą specjalnego przypadku metody znanej dziś pod nazwą metody Newtona umiano wyznaczać pierwiastki drugiego stopnia z liczb naturalnych.

Ten krótki przegląd pozwala stwierdzić, iż na przestrzeni dziejów przynajmniej kilkukrotnie opracowywano identyczne sposoby numerycznego rozwiązywania problemów matematycznych.

1.2. Czym są metody numeryczne

Metody numeryczne są poddziałem działu matematyki zwanym analizą numeryczną. Metody numeryczne, jak sama nazwa wskazuje, dotyczą głównie metod, czyli, inaczej mówiąc, algorytmów rozwiązywania w sposób przybliżony problemów występujących w nauce i technice.

Można zadać pytanie: dlaczego rozwiązania przybliżone? Przede wszystkim dlatego, że często uzyskanie rozwiązania dokładnego jest zbyt skomplikowane, zbyt długotrwałe lub wręcz niemożliwe. Niestety przybliżony charakter uzyskiwanych rozwiązań stawia przed użytkownikiem metod numerycznych wymagania dotyczące między innymi:

— ocenienia użyteczności uzyskanego rozwiązania,
— oceny dokładności tegoż rozwiązania,
— interpretacji wyników.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zazwyczaj, jeżeli nie dochodzi do sytuacji ewidentnie błędnej np. dzielenia przez zero, program zwróci jakieś wyniki, ale nie zawsze mają one sens fizyczny. Te wymagania powodują, że inżynier musi zdawać sobie sprawę z możliwych trudności i umieć upewnić się, iż rozwiązanie jest poprawne. Co prawda, istniejące systemy obliczeniowe stosowane w praktyce inżynierskiej zawierają wiele zabezpieczeń, które chronią przed popełnieniem grubych błędów. Niemniej nie zawsze domyślne wartości parametrów dają najlepsze wyniki. A jak mówi stara reguła inżynierska, są tylko trzy sposoby zapewnienia poprawnych wyników obliczeń: sprawdzenie, sprawdzenie i sprawdzenie.

1.3. Błędy metod numerycznych

Na błędy wyników obliczeń składa się cały szereg przyczyn. W literaturze najczęściej wspominane są błędy pochodzące od:

— danych wejściowych,
— metody,
— obcięcia,
— zaokrągleń,
— człowieka,
— maszyny,
— modelu obliczeniowego.

Każdy z wymienionych rodzajów błędów ma mniejszy lub większy wpływ na wynik końcowy. Niektóre z nich mają charakter błędów systematycznych, a inne błędów losowych. Niektóre można wyeliminować, a inne zależą od konstrukcji maszyn obliczeniowych i przy obecnych komputerach nie można ich uniknąć. Poniżej dokonano skrótowego omówienia wymienionych rodzajów błędów.

Błąd danych wejściowych zawsze jest obecny w obliczeniach inżynierskich, ponieważ zawsze istnieje błąd pomiaru wartości mierzonej wielkości. W związku z tym należy być wyczulonym na to zjawisko. Jest to istotne dla osób poznających warsztat inżyniera, ponieważ w zadaniach rozwiązywanych w trakcie nauki niezmiernie rzadko mówi się o niepewności danych wejściowych – wartości sił, przekrojów, mas itp. są podane jako dokładne. Z faktu istnienia tejże niepewności wynika konieczność badania stabilności algorytmów numerycznych, czyli odporności na niewielkie zakłócenia danych wejściowych. Gdy stawia się problem, należy zadać pytanie: co się stanie, gdy dane wejściowe ulegną niewielkiej zmianie? Gdy rozwiązanie jest bardzo wrażliwe na niewielkie zmiany wartości wielkości wejściowych, to mówi się o złym uwarunkowaniu problemu. Przykładem takiego źle uwarunkowanego problemu może być rozwiązanie następującego układu równań

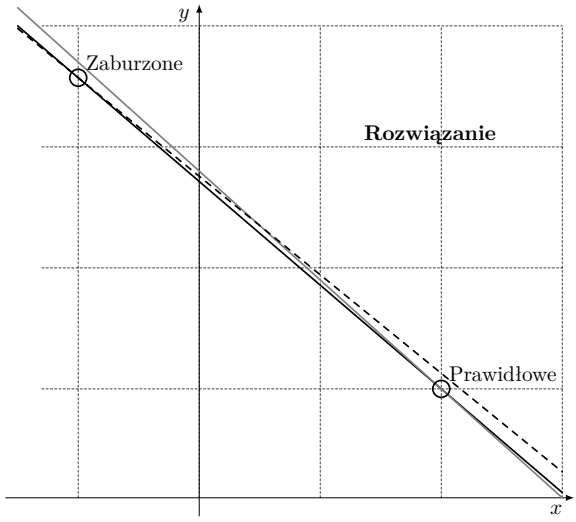
$$\begin{cases} 2x + 7y & = & 25 \\ 2{,}00001x + 7y & = & 25{,}00002 \end{cases}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązaniem tego układu jest para liczb $x = 2$ oraz $y = 3$. Gdy nastąpi niewielkie zaburzenie wartości współczynników np.

$$\begin{cases} 2x + 7y & = & 25 \\ 1{,}99999x + 7y & = & 25{,}00001 \end{cases}$$

to rozwiązaniem jest $x = -1$, $y = 3,85714$. Zatem problem jest źle uwarunkowany. Dla spostrzegawczego czytelnika takie zachowanie powyższego układu równań nie powinno być zaskoczeniem, ponieważ obydwa równania układu przedstawiają proste prawie równoległe (rysunek 1.1). Ten fakt powoduje, że niewielkie zmiany nachylenia jednej z prostych powodują duże przemieszczenia punktu przecięcia, czyli rozwiązania układu równań.



Rysunek 1.1. Przykład źle uwarunkowanego układu równań

Błąd metody wynika z natury metod numerycznych (często jest blisko powiązany z błędem obcięcia). Upraszczając opis zjawiska, pomijając składniki wyższych rzędów, czy korzystając z wybranych klas funkcji, powoduje się, że metoda nie daje wyników dokładnych. Dlatego niezwykle istotną sprawą staje się analiza błędów metody. Dzięki niej można uzyskać informacje o rzędzie wielkości tego błędu.

Błąd obcięcia jest blisko związany z błędem metody. Pojawia się wtedy, gdy opisując problem, aproksymuje się funkcje w nim występujące funkcjami z wybranej klasy. Najczęściej jest spotykany, gdy korzysta się np. z rozwinięcia funkcji w szereg Taylora i pomija (obcina) wyrazy wyższych rzędów lub gdy do obliczeń bierze się

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

skończone przedziały zamiast nieskończenie małych, na przykład przy różniczkowaniu czy całkowaniu.

Błąd zaokrągleń jest wynikiem ułomności maszyn cyfrowych, w których zamiast zbioru liczb rzeczywistych jest wykorzystywany pewien dyskretny jego podzbiór. Ten fakt z kolei powoduje konieczność zaokrąglania wyników obliczeń i czasem danych wejściowych do liczb z tegoż zbioru. Gdy komputer posługuje się arytmetyką binarną, konieczne jest też dokonanie konwersji z systemu dziesiętnego do binarnego, co często niesie za sobą konieczność zaokrąglenia konwertowanej liczby.

Standard IEEE dotyczący zmiennoprzecinkowej arytmetyki (IEEE 754) w swej aktualnej wersji (IEEE 754-2008) określa sposób kodowania liczb zmiennoprzecinkowych (rzeczywistych) w systemach cyfrowych. Standard określa między innym następujące zagadnienia:

— formaty arytmetyczne: binarne i decymalne dane zmiennoprzecinkowe o skończonej precyzji, włączając w to specjalne wartości (zero ujemne, nieskończoności oraz NaN – nie liczba),
— formaty wymiany danych: kodowanie pozwalające na wyminę danych zmiennoprzecinkowych między różnymi systemami,
— algorytmy zaokrąglania – używane do zaokrąglania w czasie obliczeń i konwersji liczb,
— operacje arytmetyczne i inne operacje na formatach arytmetycznych,
— obsługę sytuacji wyjątkowych – takich jak: dzielenie przez zero, logarytmowanie liczb ujemnych, przepełnienia (przekroczenia dostępnego zakresu liczbowego).

Najistotniejsza dla osoby zajmującej się metodami numerycznymi część tego unormowania dotyczy binarnej reprezentacji liczb rzeczywistych. Dla przykładu w tabeli 1.1 przedstawiono dwa typy liczb określanych jako binary64 oraz binary32 (typy double oraz single w języku C). Zazwyczaj liczby zmiennoprzecinkowe są pamiętane w postaci struktury, w której kolejno od lewej występują: bit znaku, wykładnik (cecha) oraz mantysa.

Mantysa jest zapisywana w postaci liczby w formacie stałoprzecinkowym. Ponieważ liczba zawsze jest normalizowana tak, by najwyższy bit był ustawiony, nie zapisuje się go, co daje jeden bit więcej na część ułamkową. Długość pola przeznaczonego na mantysę decyduje o dokładności reprezentacji liczb rzeczywistych przez zmiennoprzecinkowe i tak, jeżeli są 52 bity części ułamkowej (typ double w języku C) – najmniejszą różnicą między dwoma liczbami, jaką można zapisać, jest 2^{-52}, czyli około $2,2 \cdot 10^{-16}$. W tak zapisanej liczbie 15 cyfr jest znaczących. Z kolei dla liczb zapisanych w zmiennych pojedynczej precyzji (długość 32 bitów), gdzie 23 bity opisują część ułamkową mantysy, najmniejsza różnica między liczbami to ok. $1,2 \cdot 10^{-7}$, zapisuje się więc 6 cyfr znaczących.

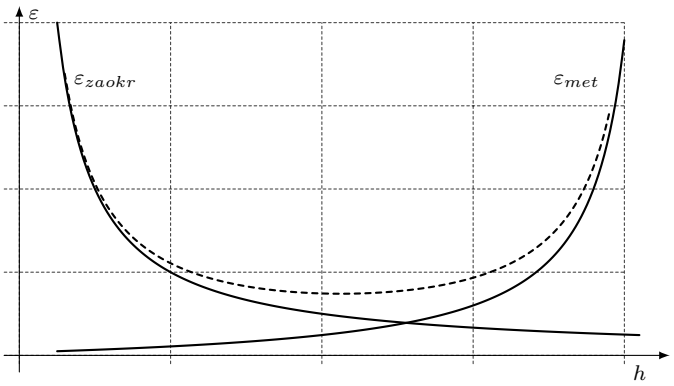
W obliczeniach bardzo istotna jest długość części przeznaczonej na mantysę. Na tej podstawie wyznacza się precyzję arytmetyki komputera. Jeżeli mantysa jest zapisywana za pomocą $t = 52$ bitów, to precyzja wynosi $\varepsilon = 2^{-t-1} = 2^{-53}$. Wielkość precyzji w zgrubny sposób określa wielkość błędów konwersji i zaokrąglenia. Warto pamiętać, iż działania na liczbach zmiennoprzecinkowych nie są łączne ani rozdzielne.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 1.1. Schemat pól bitowych formatu zmiennoprzecinkowego 32- i 64-bitowego

	Długość [bit]	
Pole	binary64	binary32
Znak	1	1
Wykładnik (cecha)	11	8
Mantysa	52	23
Bity (razem)	64	32
Przesunięcie wykładnika	1023	127
Bity precyzji	53	24

To oznacza, że kolejność wykonywania działań ma znaczenie i może znacząco wpłynąć na wynik.



Rysunek 1.2. Zmiany błędu metody i zaokrągleń w funkcji kroku

W obliczeniach numerycznych błędy występują równocześnie. Z powodu takiej sytuacji trudno jest dać zalecenia co do wyboru parametrów obliczeń. Rysunek 1.2 wyjaśnia taką sytuację. Przedstawiono na nim błąd zaokrągleń i błąd metody przy całkowaniu kwadraturą Newtona–Cotesa. Wraz ze wzrostem długości kroku maleją względne błędy zaokrągleń, ale rośnie błąd metody. Sumaryczny błąd – zaznaczony linią przerywaną – wykazuje minimum dla pewnej długości kroku całkowania.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Błąd człowieka wynika z pogody, stanu psychofizycznego, braku wiedzy, nieuwagi i innych czynników, a sprowadza się do pomyłek.

Błąd maszyny ma charakter katastroficzny i świadczy o usterce w sprzęcie lub oprogramowaniu. W obliczeniach wykonywanych za pomocą komputerów domowych może przysporzyć sporo problemów, jako trudny do zdiagnozowania. Może się pojawić, gdy ktoś usiłuje podnosić osiągi komputera przez zmianę wartości częstotliwości pracy podzespołów powyżej zalecanych przez producenta.

Błąd modelu obliczeniowego nie jest bezpośrednio związany z metodami numerycznymi, niemniej może być źródłem poważnych problemów, gdy okaże się, że rozwiązywano problem całkowicie odmienny od postawionego. Może się zdarzyć, iż zamiast badania odpowiedzi dynamicznej układu będzie analizowany problem statyczny.

Innym przykładem błędnego modelu obliczeniowego może być próba analizowania procesów przeróbki plastycznej (np. metodą elementów skończonych) przy modelu materiału, który nie uwzględnia zjawiska umocnienia materiału.

Miary błędów. Błąd można określać jako wartość bezwzględną różnicy między wartością przybliżoną $\hat{x}$ a wartością dokładną x. Tak obliczony błąd nazywa się błędem bezwzględnym

$$\Delta x = |\hat{x} - x| \tag{1.1}$$

Ponieważ obliczając wielkość tego błędu, używa się wartości bezwzględnej, nie jest istotna kolejność odejmowania. Odnosząc wartość błędu bezwzględnego do wartości uznawanej za dokładną, uzyskuje się błąd względny

$$\varepsilon = \frac{|\hat{x} - x|}{x} = \frac{\Delta x}{x} \tag{1.2}$$

Zaokrąglanie i obcinanie. Jak już zostało powiedziane w poprzednim podrozdziale, maszyny cyfrowe nie posługują się liczbami rzeczywistymi, a jedynie pewnym ich podzbiorem. Wielkość tego podzbioru zależy od sposobu reprezentacji liczb w systemie. Jeszcze kilkanaście lat temu często można było spotkać informację, iż dane obliczenia należy wykonać w podwójnej precyzji. W tej chwili wszystkie pakiety obliczeniowe posługują się tak zapisywanymi liczbami. Pojawiają się narzędzia (biblioteki procedur obliczeniowych) wykorzystujące do obliczeń wydajne procesory kart graficznych, posiadające rejestry o długości 80 i więcej bitów. W związku z tym w niektórych językach programowania można znaleźć typy danych pozwalające na zapisywanie więcej niż 15 cyfr znaczących. Zawsze jednak zachodzi zjawisko zaokrąglania lub obcinania. Obcinanie jest sposobem zaokrąglania zawsze w kierunku zera. Niech $\tilde{x}$ oznacza maszynową reprezentację liczby rzeczywistej x zgodnie z równaniem $\tilde{x} = fl(x)$, gdzie $fl(\cdot)$ oznacza konwersję liczby rzeczywistej do jej maszynowego odpowiednika. Błąd względny zaokrąglenia przy obcinaniu spełnia warunek

$$\delta = \left| \frac{x - \tilde{x}}{x} \right| \leqslant 2^{-t}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tradycyjne zaokrąglanie przeprowadza się w kierunku zera, gdy następna pozycja po zaokrąglanej zawiera cyfry 0, 1, 2, 3 lub 4, zaś w kierunku nieskończoności (o znaku zaokrąglanej liczby) w pozostałych przypadkach. W ten sposób błąd popełniany przy zaokrąglaniu spełnia warunek

$$\delta = \left| \frac{x - \tilde{x}}{x} \right| \leqslant \frac{1}{2} 2^{-t} = 2^{-t-1} = \varepsilon$$

Widać, że obcięcie wprowadza dwa razy większy błąd niż zaokrąglanie. Omawiając te zagadnienia, nie sposób nie wspomnieć o pojęciach precyzji i dokładności. Pod pojęciem dokładności rozumie się zazwyczaj określenie wielkości błędu popełnianego przy obliczeniach. Precyzja wiąże się z reprezentacją liczb rzeczywistych w maszynach cyfrowych i określa liczbę cyfr używanych do obliczeń.

Można ogólnie przyjąć, iż maksymalny błąd popełniany przy zaokrąglaniu liczb rzeczywistych nie będzie większy niż precyzja (ε), czyli

$$fl(x) = x(1 + \delta), \quad \text{gdzie } |\delta| \leqslant \varepsilon$$

Powyższa zależność pozwala na oszacowanie wielkości błędu zaokrągleń popełnianego przy poszczególnych operacjach arytmetycznych. Można przyjąć, iż w trakcie wykonywania obliczeń numerycznych jest prawdziwa następująca zależność na wielkość błędu wyniku działań arytmetycznych oznaczonych przez $\odot$

$$fl(x \odot y) = (x \odot y)(1 + \delta), \text{ gdzie } \delta \leqslant \varepsilon$$

Warto jeszcze wspomnieć, że przy wykonywaniu wielu operacji błędy narastają. Na przykład przy sumowaniu n liczb błąd względny sumy będzie równy około $n\varepsilon$.

Bardzo ważnym działaniem występującym chyba w każdej metodzie numerycznej jest odejmowanie. Często odejmowane są bliskie sobie wartości. W takim przypadku dochodzi do sytuacji utraty cyfr znaczących. Można to przeanalizować na przykładzie odejmowania dwóch wartości przy pięciocyfrowej precyzji. Niech pośrednie wyniki obliczeń, zapamiętane z precyzją siedmiocyfrową wynoszą $x_1 = 1,7459082$, $x_2 = 1,7458231$. Różnica wyniesie $x_1 - x_2 = 0,0000851$, co zostanie zapamiętane jako $fl(x_1 - x_2) = 0,851 \cdot 10^{-4}$, liczba ta ma tylko 3 cyfry znaczące. Więc nawet gdy obliczenia wewnątrz procesora są wykonywane z większą precyzją niż liczby zapisane w pamięci, skończona długość mantysy powoduje, że następuje utrata cyfr znaczących.

Przenoszenie błędów jest zjawiskiem spotykanym na co dzień w praktyce obliczeń numerycznych. Niestety, nie jest łatwo określić wpływ poszczególnych źródeł błędów na końcowy wynik obliczeń numerycznych. Jest to tym trudniejsze, iż często wynik końcowy otrzymuje się po wielu działaniach wykonywanych szeregowo. Dlatego najłatwiej jest wykorzystać poniższą zależność, określającą „wrażliwość" wyniku końcowego na zaburzenia poszczególnych wielkości wchodzących do obliczeń. Wielkość błędu bezwzględnego dla zależności $y = f(x_i)$, gdzie x_i oznacza poszczególne wielkości wejściowe do obliczeń, można określić następująco

$$\Delta y \approx \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} \Delta x_i \tag{1.3}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

skąd uzyskuje się oszacowanie wielkości błędu maksymalnego w postaci

$$|\Delta y| \leqslant \left| \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} \right| |\Delta x_i| \tag{1.4}$$

W zależności (1.4) dokonuje się aproksymacji liniowej zależności $f(x)$ w otoczeniu wartości dokładnej. Omawiana zależność tylko w przybliżony sposób pozwala na określenie wielkości błędu. Ponieważ poszczególne pochodne cząstkowe w otoczeniu wartości dokładnej mogą ulegać znacznym zmianom, należy do analizy błędów wziąć maksima wartości bezwzględnych pochodnych cząstkowych w otoczeniu wartości dokładnej.

Przedstawione zależności prowadzą do bardzo pesymistycznego oszacowania wielkości popełnianego błędu. Zakłada się najgorszą z możliwych konfiguracji. W związku z tym mówi się też o niepewności obliczeń i bierze pod uwagę statystyczną charakterystykę błędu – określając, w jakim zakresie występuje odpowiedni odsetek wyników (95%, 99%).

1.4. Poruszane problemy

Podstawowym celem niniejszego skryptu jest w miarę przystępne ujęcie klasycznych zagadnień metod numerycznych. Nie przedstawiano dowodów twierdzeń czy skomplikowanych wyprowadzeń. Zawarte zostały jedynie podstawowe wyprowadzenia, niezbędne do rozumienia wykładu.

W ramach kursu omawiane będą metody dotyczące:

— rozwiązywania równań nieliniowych,
— wyznaczania wartości całek oznaczonych,
— różniczkowania numerycznego,
— rozwiązywania równań różniczkowych zwyczajnych i cząstkowych,
— rozwiązywania układów równań liniowych,
— aproksymacji i interpolacji,
— oraz planowania eksperymentu.

Oprócz tych zagadnień omawianych w ramach przedmiotu *Metody obliczeniowe i planowanie eksperymentu*, poruszono zagadnienia nieco bardziej skomplikowane jak podstawy metody elementów skończonych czy metody elementów brzegowych.

W ramach rozszerzenia materiału teoretycznego przedstawiono wprowadzenie do pakietów SageMath, Maxima oraz DAKOTA, których znajomość może ułatwić pracę inżyniera. Autorzy dokładali starań, by przedstawiane przykłady miały, przynajmniej w części, charakter praktyczny.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 2

Interpolacja

Problem interpolacji można sformułować jako poszukiwanie funkcji przechodzącej przez zadane punkty, których odcięte zwane są węzłami interpolacji. Innymi słowy, poszukiwane są wartości przyjmowane przez funkcję pomiędzy węzłami. Ponieważ problemy inżynierskie najczęściej rozwiązuje się metodami przybliżonymi, polegającymi na podziale analizowanego obszaru na podobszary, otrzymuje się wyniki w dyskretnych punktach. Między innymi z tego powodu interpolacja jest jednym z najczęściej napotykanych problemów w dzisiejszej nauce i technice. Poza tym, metoda elementów skończonych czy metoda elementów brzegowych oraz wiele innych metod numerycznych wykorzystują interpolację nieznanej funkcji, stanowiącej rozwiązanie problemu, do znalezienia przybliżonego rozwiązania.

Okazuje się też, że zdecydowana większość informacji przez wieki utrwalanej przez ludzkość miała charakter dyskretny, zaś w ostatnich latach, na skutek wprowadzenia techniki cyfrowej, to zjawisko uległo jeszcze większemu nasileniu. A zatem, ponieważ informacje są przechowywane w postaci dyskretnej, niezbędne jest do prezentacji tej informacji użycie interpolacji.

Interpolacja oznacza znalezienie funkcji $Q(x)$ takiej, że

$$Q(x_i) = y_i \tag{2.1}$$

dla zbioru węzłów x_i z wartościami y_i dla $i = 0, 1, 2, \ldots, n$. Przy interpolacji podejmuje się arbitralną decyzję dotyczącą klasy funkcji, wśród których poszukuje się funkcji $Q(x)$. Zbiór, z którego są wybierane funkcje interpolujące $q_j(x)$, nazywa się bazą interpolacji.

$$Q(x) = \sum_{j=0}^{N} a_j q_j(x) \tag{2.2}$$

W przypadkach spotykanych w praktyce inżynierskiej przyjęta baza interpolacji może zależeć od zjawiska fizycznego związanego z interpolowaną funkcją. Bardzo dużą gamę procesów można dobrze interpolować wielomianami, jednak zjawiska o charakterze okresowym znacznie lepiej daje się odwzorować za pomocą funkcji okresowych jak sinus, czy cosinus.

Zatem jako bazę interpolacji można zastosować jednomiany 1, x, x^2, $\ldots$, lub funkcje trygonometryczne $(1, \sin x, \sin 2x, \ldots)$ czy funkcje wykładnicze $(1, e^x, \ldots)$. W zależności od przyjętej bazy mówi się o różnych rodzajach interpolacji. Zagadnienie wyboru bazy interpolacji jest bardzo szerokie i wykracza poza zakres niniejszego opracowania. Dalsze rozważania zostaną ograniczone do bazy wielomianowej, a ści-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

ślej rzecz biorąc, do jednomianów z krótkim wtrąceniem dotyczącym interpolacji za pomocą wielomianów Czebyszewa.

2.1. Interpolacja wielomianowa

Pojęcie interpolacji wielomianowej można zdefiniować jako poszukiwanie wielomianu $W(x)$ najniższego stopnia, takiego że dla danych $n+1$ węzłów x_i, w których są dane wartości y_i, zachodzi równość

$$W(x_i) = y_i, \quad \text{dla} \quad i = 0, 1, 2, \ldots, n \tag{2.3}$$

Tak sformułowane zadanie można rozwiązywać na różne sposoby, niektóre z nich nadają się bardziej do obliczeń prowadzonych na papierze, inne z kolei zapewniają minimalizację wpływu błędów zaokrągleń na wynik końcowy.

Zapewne dla wielu czytelników naturalnym sposobem rozwiązania problemu interpolacji jest zbudowanie układu równań, pozwalającego na wyznaczenie wartości współczynników wielomianu interpolacyjnego. Aby skonstruować taki układ równań, można wykorzystać wprost równanie (2.2). Liczba węzłów wyznacza maksymalny stopień wielomianu, który będzie użyty do interpolacji

$$W(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

Zatem nieznane pozostają współczynniki a_i, które należy wyznaczyć, aby uzyskać pełną postać wielomianu interpolacyjnego. Można tego dokonać, zapisując dla każdego węzła równanie w postaci (2.2), co daje układ równań liniowych

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n &= y_1 \\ a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_n x_2^n &= y_2 \\ \qquad\qquad \vdots & \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n &= y_n \end{cases} \tag{2.4}$$

Taki układ równań można zapisać w postaci macierzowej

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}}_{\mathbb{A}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbb{Y}} \tag{2.5}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązanie układu równań (2.5) pozwala na wyznaczenie współczynników wielomianu interpolacyjnego $W(x)$. Warto przy tej okazji wspomnieć, iż macierz $\mathbb{X}$ nazywa się macierzą Vandermonde'a.

Niemniej metoda taka ma znaczące wady. Przede wszystkim nie jest to metoda wydajna obliczeniowo, oprócz tego często macierz $\mathbb{X}$ staje się źle uwarunkowana, co prowadzi do dużych błędów tak wyznaczonych współczynników wielomianu. Macierz $\mathbb{X}$ staje się źle uwarunkowana najczęściej wraz ze wzrostem liczby węzłów interpolacji.

Pewnym ulepszeniem tej metody jest zastosowanie bazy w postaci wielomianów Czebyszewa, zamiast jednomianów $1, x, x^2, \ldots$ Uzyskuje się wtedy lepiej uwarunkowaną macierz $\mathbb{X}$. Ale metoda dalej pozostaje bardzo nieekonomiczna obliczeniowo.

2.1.1. Interpolacja wielomianami Czebyszewa

Wielomiany Czebyszewa pierwszego rodzaju można definiować na różne sposoby. Rekurencyjna definicja tych wielomianów jest następująca

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \end{aligned} \tag{2.6}$$

Na podstawie tak sformułowanej definicji łatwo można wyznaczyć postać poszczególnych wielomianów Czebyszewa pierwszego rodzaju; wielomian podstawowy i cztery kolejne wyrażają się zależnościami

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \end{aligned} \tag{2.7}$$

Wielomiany Czebyszewa mają wiele ciekawych właściwości [1]. Jedną z istotniejszych, z punktu widzenia osób zajmujących się metodami numerycznymi, jest fakt, iż spośród wszystkich wielomianów stopnia n mających współczynnik równy 1 przy najwyższej potędze, wielomian Czebyszewa

$$\frac{1}{2^{n-1}} T_n(x)$$

w przedziale $\langle -1, 1 \rangle$ ma najmniejsze oscylacje, czyli najmniejsze odchylenie od zera.

Inną istotną właściwością tych wielomianów jest to, że można określić ogólną postać zależności na położenie n miejsc zerowych wielomianu Czebyszewa $T_n(x)$ w postaci

$$x_i = \cos\left(\frac{2i+1}{2n}\pi\right), \quad \text{dla } i = 1, 2, \ldots, n \tag{2.8}$$

Warto też pamiętać, że wybór węzłów interpolacji w miejscach, gdzie wielomiany Czebyszewa zerują się, daje w praktyce wiele korzyści. Przede wszystkim taki wybór węzłów zmniejsza oscylacje wynikowego wielomianu interpolującego. Należy pamiętać, że

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

o wielomianach Czebyszewa zazwyczaj mówi się w odniesieniu do przedziału $\langle -1, 1\rangle$, wobec tego przy wykorzystywaniu ich miejsc zerowych do interpolacji konieczne jest przekształcenie przedziału $\langle a, b\rangle$ w przedział $\langle -1, 1\rangle$ za pomocą poniższych zależności

$$\begin{aligned} x_{\langle a,\,b\rangle} &= \frac{a+b}{2} + \frac{b-a}{2} x_{\langle -1,\,1\rangle} \\ x_{\langle -1,\,1\rangle} &= \frac{2}{b-1}\left(x_{\langle a,\,b\rangle} - \frac{a+b}{2} \right) \end{aligned} \tag{2.9}$$

gdzie $x_{\langle -1,\,1\rangle}$ jest wartością x z przedziału $\langle -1, 1\rangle$, zaś $x_{\langle a,\,b\rangle}$ jest jej odpowiednikiem w przedziale $\langle a, b\rangle$.

Gdy zastosowana zostanie baza w postaci wielomianów Czebyszewa zamiast jednomianów $1, x, x^2, \ldots$, układ równań do interpolacji przyjmie następującą postać

$$\begin{bmatrix} T_0(x_0) & T_1(x_0) & T_2(x_0) & \ldots & T_n(x_0) \\ T_0(x_1) & T_1(x_1) & T_2(x_1) & \ldots & T_n(x_1) \\ T_0(x_2) & T_1(x_2) & T_2(x_2) & \ldots & T_n(x_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_0(x_n) & T_1(x_n) & T_2(x_n) & \ldots & T_n(x_n) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{2.10}$$

Po rozwiązaniu układu równań (2.10) uzyskuje się współczynniki c_i stojące przy kolejnych wielomianach Czebyszewa i w ten sposób wyznacza się postać odpowiedniego wielomianu interpolującego.

$$W(x) = c_0 T_0(x) + c_1 T_1(x) + \ldots + c_n T_n(x)$$

Wyniki powinny być identyczne z rozwiązaniem problemu interpolacji poprzednią metodą. Jednakże zysk polega na dużo lepszym uwarunkowaniu układu (2.10) niż układu (2.5).

Szczególnie korzystne jest przyjęcie węzłów w miejscach zerowych wielomianu Czebyszewa oraz nieco zmodyfikowanej bazy wielomianowej, tak by pierwszy wielomian T_0 był równy $1/\sqrt{2}$. Po takim zabiegu szczególnie łatwo jest odwrócić uzyskaną macierz Vandermonde'a

$$\mathbb{X}^{-1} = \frac{2}{n+1} \mathbb{X}^T$$

Zaś wskaźnik uwarunkowania macierzy $\mathbb{X}$ jest bliski jedności, co oznacza, że błędy zaokrągleń przy odwracaniu macierzy będą miały niewielki wpływ na wynik końcowy.

Żyjący w latach 1821 – 1894 rosyjski matematyk Pafnutij Lwowicz Czebyszew miał znaczący wkład w rozwój matematyki. Rozwiązując problem optymalizacji mechanizmu maszyny parowej, wyznaczył rodzinę wielomianów ortogonalnych, które są często wykorzystywane w analizie numerycznej i nie tylko.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

2.1.2. Przykład obliczeniowy

Sformułowanie problemu. Wyznaczyć współczynniki wielomianu interpolującego funkcję $y = \sin x$ w czterech równoodległych punktach.

Rozwiązanie. Poniżej zestawiono współrzędne węzłów interpolacji i wartości funkcji $\sin x$

x	0	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
y	0	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	1

Ponieważ cztery węzły pozwalają na wyznaczenie czterech współczynników wielomianu, układ równań (2.5) przyjmie następującą postać

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0{,}52360 & 0{,}27416 & 0{,}14355 \\ 1 & 1{,}04720 & 1{,}09662 & 1{,}14838 \\ 1 & 1{,}57080 & 2{,}46740 & 3{,}87578 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1{,}02043 \\ -0{,}06547 \\ -0{,}11387 \end{bmatrix}$$

Rozwiązując ten układ równań, można wyznaczyć poszukiwane współczynniki a_i, zaś wynikowy wielomian ma postać

$$W(x) = -0{,}11387x^3 - 0{,}06547x^2 + 1{,}02043x$$

Identyczny wynik otrzyma się, gdy do interpolacji zastosuje się wielomiany Czebyszewa. Nie powinno być to zaskoczeniem, ponieważ jedno z ważniejszych twierdzeń algebry mówi, iż dla danego zbioru węzłów istnieje tylko jeden wielomian stanowiący rozwiązanie problemu interpolacji [23].

2.1.3. Interpolacja Newtona

Już wiele lat temu zostały opracowane znacznie wydajniejsze metody wyznaczania wielomianu interpolacyjnego. Jedną z nich jest interpolacja Newtona. W metodzie interpolacji Newtona wielomian interpolacyjny wyznacza się na podstawie następującej zależności

$$W(x) = \sum_{i=0}^{k} c_i \prod_{j=0}^{i-1} (x - x_j) \tag{2.11}$$

Czasem łatwiej można zapamiętać ten wzór w postaci rozwiniętej

$$W(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_2) + \dots \tag{2.12}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Interpolacja Newtona wykorzystuje do interpolacji rodzinę trójkątną wielomianów konstruowanych jako iloczyny dwumianów. W ten sposób problem interpolacji sprowadzony został do problemu wyznaczenia wartości współczynników c_i. Najwygodniej można je wyznaczyć, posługując się ilorazami różnicowymi kolejnych rzędów. Tworzenie tych ilorazów zostanie teraz pokrótce omówione.

Ilorazy różnicowe, oznaczane przez $f[x]$, definiowane są w następujący sposób, w zależności od rzędu:

— iloraz różnicowy rzędu zerowego

$$f[x_0] = f(x_0),$$

— iloraz różnicowy rzędu pierwszego

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

— iloraz różnicowy rzędu drugiego

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0},$$

— ilorazy różnicowe kolejnych rzędów wyznacza się analogicznie, w liczniku znajduje się różnica ilorazów różnicowych niższego rzędu, zaś w mianowniku jest różnica skrajnych wartości argumentów z tych ilorazów różnicowych.

Ilorazy różnicowe zapisuje się w postaci tablicy. Kolejne kolumny zawierają ilorazy różnicowe coraz wyższych rzędów. Pierwsza i druga kolumna zawierają odpowiednio wartości współrzędnych x_i oraz $y_i = f[x_i]$ kolejnych węzłów. Aby uzyskać dużą dokładność, w obliczeniach maszynowych dobrze jest ułożyć węzły w kolejności rosnącej lub malejącej względem x.

x	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
x_0	y_0	$\frac{y_1-y_0}{x_1-x_0} = f_{01}$	$\frac{f_{12}-f_{01}}{x_2-x_0} = f_{012}$	$\frac{f_{123}-f_{012}}{x_3-x_0}$
x_1	y_1	$\frac{y_2-y_1}{x_2-x_1} = f_{12}$	$\frac{f_{23}-f_{12}}{x_3-x_1} = f_{123}$	$\vdots$
x_2	y_2	$\frac{y_3-y_2}{x_3-x_2} = f_{23}$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$		

Jak widać, tablica ilorazów różnicowych ma kształt trójkąta. Liczba wyliczonych kolumn ilorazów różnicowych zależy od wartości współrzędnych węzłów. W skrajnym przypadku tych kolumn może być o jedną mniej niż węzłów interpolacji. Ilorazy różnicowe w pierwszym wierszu tablicy stanowią współczynniki c_i wielomianu Newtona.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

2.1.4. Przykład obliczeniowy

Sformułowanie problemu. Dokonać interpolacji wielomianowej funkcji $\sin x$ za pomocą wzoru Newtona w przedziale $\langle 0, \pi/2 \rangle$, wybierając cztery równoodległe węzły, czyli realizując podział na trzy równe przedziały. Określić błąd interpolacji.

Rozwiązanie. Korzystając ze wzoru Newtona, otrzymuje się następującą postać wielomianu

$$W(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_2) + \cdots$$

Aby wyznaczyć poszukiwane współczynniki, należy zbudować tablicę ilorazów różnicowych dla punktów stanowiących węzły interpolacji, tak jak poniżej

x	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, \ldots, x_{i+3}]$
0,00000	0,00000	0,95493	−0,24434	−0,11387
0,52360	0,50000	0,69906	−0,42321	
1,04720	0,86603	0,25587		
1,57080	1,00000			

Współczynniki wielomianu interpolującego przyjmują następujące wartości

c_0	c_1	c_2	c_3
0	0,95493	−0,24434	−0,11387

Zatem wielomian interpolacyjny uzyskuje następującą postać

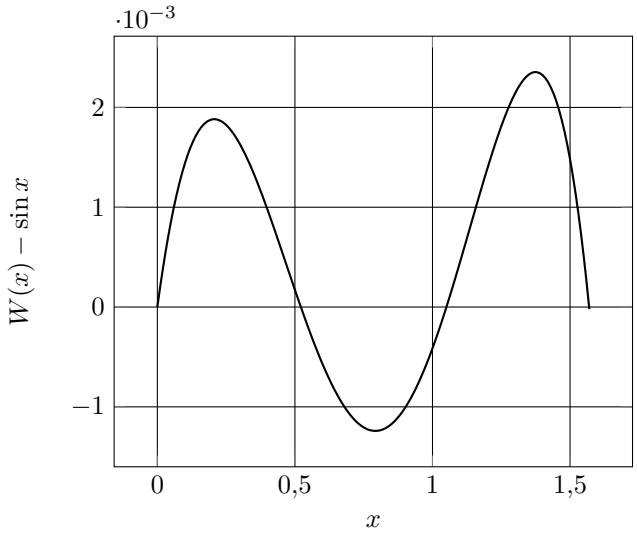
$$W(x) = 0 + 0{,}95493(x - 0) - 0{,}24434(x - 0)(x - 0{,}5236) - \\ - 0{,}11387(x - 0)(x - 0{,}5236)(x - 1{,}0472)$$

Po niezbędnych uproszczeniach wielomian można zapisać jako

$$W(x) = 1{,}0204x - 0{,}065473x^2 - 0{,}11387x^3$$

Określenie wielkości błędu w wybranych punktach przedziału interpolacji można przeprowadzić, porównując wartości uzyskanego wielomianu z wartościami funkcji $\sin x$, co zostało przedstawione na rysunku 2.1.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 2.1. Błąd interpolacji funkcji $y = sin(x)$ za pomocą wielomianu stopnia trzeciego z węzłami równoodległymi

2.1.5. Przykład obliczeniowy

Sformułowanie problemu. Porównać wynik interpolacji funkcji $y = \sin x$ z poprzedniego przykładu z węzłami równoodległymi z interpolacją przeprowadzoną przy węzłach nierównoodległych, wybranych w miejscach zerowych wielomianu Czebyszewa czwartego stopnia.

Rozwiązanie. Miejsca zerowe wielomianów Czebyszewa stopnia n wyrażają się wzorem (2.8).

$$x_i = \cos\left(\frac{2i+1}{2n}\pi\right), \quad \text{dla } i = 1, 2, \ldots, n$$

Wielomiany Czebyszewa mają miejsca zerowe x_c w przedziale $\langle -1, 1\rangle$, zatem konieczne jest przekształcenie przedziału $\langle 0, \pi/2\rangle$ do powyższego, uzyskuje się to za pomocą transformacji (2.9).

$$x_n = \frac{a+b}{2} + x_c\frac{b-a}{2}$$

Wielomian Czebyszewa czwartego stopnia przedstawia się następującą zależnością

$$T_4(x) = 8x^4 - 8x^2 + 1$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Miejsca zerowe x_c wielomianu Czebyszewa $T_4(x)$, oraz odpowiadające im wartości x_n w przedziale $\langle 0, \pi/2 \rangle$, wyznaczone na podstawie zależności (2.9) uporządkowane w kolejności malejącej, są następujące

x_c	0,92388	0,38268	−0,38268	−0,92388
x_n	1,511011	1,085957	0,484839	0,059785

Następnie, aby określić współczynniki niezbędne do zbudowania wielomianu Newtona, należy zbudować tablicę ilorazów różnicowych jak poniżej

x	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, \ldots, x_{i+3}]$
1,51101	0,99821	0,26694	−0,41861	−0,11426
1,08596	0,88475	0,69651	−0,25279	
0,48484	0,46607	0,95592		
0,05978	0,05975			

Po odczytaniu z powyższej tablicy wartości współczynników c_i, wielomian wynikowy można zapisać jako

$$W(x) = 0{,}99821 + 0{,}26694(x - 1{,}51101) - 0{,}41861(x - 1{,}51101)(x - 1{,}08596) - \\ - 0{,}11426(x - 1{,}51101)(x - 1{,}08596)(x - 0{,}48484)$$

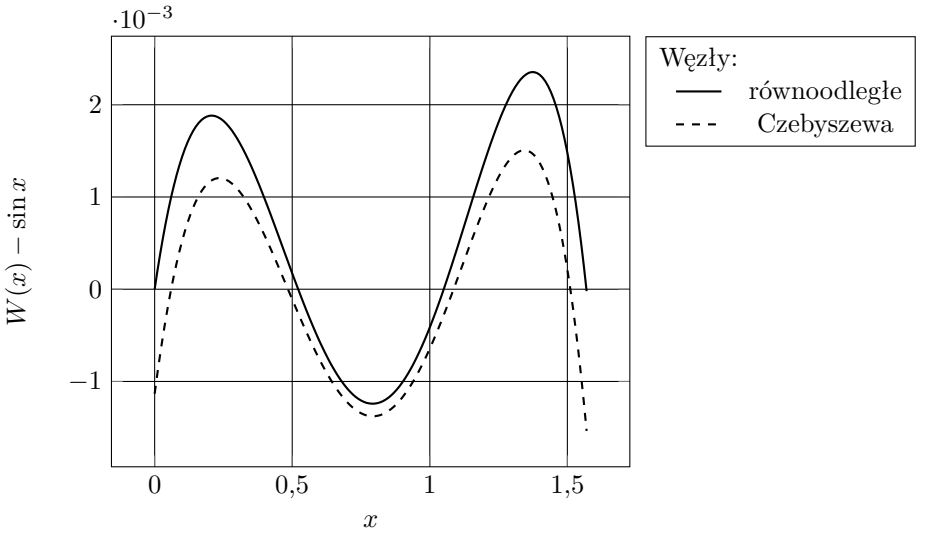
Po niezbędnych przekształceniach można go zapiać w następującej postaci

$$W(x) = -1{,}13258 + 1{,}02270x - 0{,}06648x^2 - 0{,}11426x^3$$

Porównanie wartości błędów dla interpolacji z węzłami równoodległymi i z węzłami Czebyszewa jest przedstawione na rysunku 2.2. Jak widać, przy węzłach Czebyszewa wielomian ma mniejszą amplitudę zmian błędu interpolacji. Węzły interpolacji w jednym i drugim przypadku znajdują się w miejscach, gdzie wykres błędu przyjmuje wartość zero. Interpolacja wielomianowa z węzłami Czebyszewa daje więc większe błędy na krańcach przedziału interpolacji, za to obserwuje się zdecydowanie mniejsze oscylacje przebiegu błędu bezwzględnego interpolacji. Pojawienie się błędów na krańcach przedziału jest naturalne, ponieważ nie ma węzłów interpolacji na tychże krańcach.

Czy można uznać, iż uzyskany wielomian dosyć dobrze przedstawia przebieg funkcji $y = \sin x$? Odpowiedź na takie pytanie silnie zależy od celu przeprowadzanej interpolacji. Gdyby w ten sposób miał być generowany na przykład przebieg elektryczny lub akustyczny, należałoby przeprowadzić dalsze testy, które pozwoliłyby określić jego przydatność do konkretnego zastosowania.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 2.2. Błąd interpolacji funkcji $y = sin(x)$ za pomocą wielomianu stopnia trzeciego z węzłami Czebyszewa i węzłami równoodległymi

Niemniej, uzyskano wielomian mający trzy współczynniki, pozwalający uzyskać przebieg funkcji sinus z dokładnością na poziomie pojedynczych procentów. Uzyskany wielomian dobrze opisuje funkcję $y = \sin x$ tylko dla $x \in \langle 0, \pi/2 \rangle$. Zatem oczywiste powinno być, że próba skonstruowania dłuższego przebiegu sinusoidalnego z otrzymanego wielomianu, wymaga „składania" wynikowej funkcji sinus z ćwiartek opisywanych wielomianem.

2.1.6. Interpolacja Lagrange'a

Interpolacja Lagrange'a przebiega w nieco inny sposób niż interpolacja Newtona. Zależności pozwalające wyznaczyć wielomian najmniejszego stopnia, przechodzący przez wszystkie wartości w węzłach interpolacji wyglądają następująco

$$p(x) = \sum_{i=0}^{n} y_i l_i, \quad \text{gdzie} \quad l_i = \prod_{j=0,\, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{2.13}$$

Zatem wynikowy wielomian interpolacyjny jest kombinacją liniową wartości funkcji interpolowanej w węzłach interpolacji i wielomianów Lagrange'a, zapisanych dla danego węzła.

Łatwo sprawdzić, że wielomiany Lagrange'a przyjmują wartość jeden dla jednego węzła, w pozostałych zaś węzłach mają miejsca zerowe. To właśnie wielomiany Lagrange'a są często wykorzystywane w metodach przybliżonego rozwiązywania równań różniczkowych na przykład metodzie elementów skończonych czy metodzie elementów brzegowych. Ta cecha tych wielomianów ma duże znaczenie przy fizycznej interpretacji i wyprowadzaniu równań elementów skończonych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

2.1.7. Przykład obliczeniowy

Sformułowanie problemu. Posługując się wzorem Lagrange'a, wyznaczyć wielomian interpolacyjny dla punktów o współrzędnych podanych poniżej

i	x	y
0	-2	$-0{,}6$
1	-1	0,8
2	0	1,0
3	1	1,2
4	2	2,6

Rozwiązanie. Korzystając z zależności (2.13), można zapisać wielomiany Lagrange'a dla kolejnych węzłów

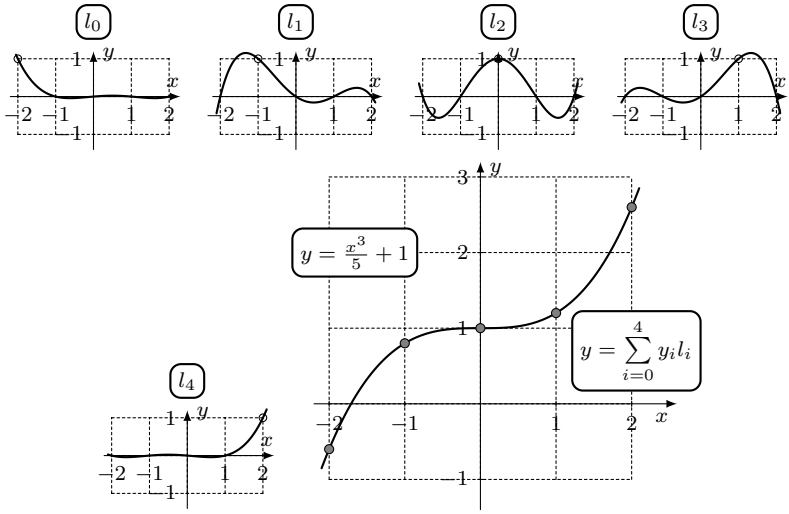
$$\begin{aligned} l_0 &= \frac{x\ (x-1)\ (x+1)\ (x-2)}{24} \\ l_1 &= -\frac{x\ (x-1)\ (x-2)\ (x+2)}{6} \\ l_2 &= \frac{(x-1)\ (x+1)\ (x-2)\ (x+2)}{4} \\ l_3 &= -\frac{x\ (x+1)\ (x-2)\ (x+2)}{6} \\ l_4 &= \frac{x\ (x-1)\ (x+1)\ (x+2)}{24} \end{aligned}$$

W następnym kroku należy wyznaczyć końcową postać wielomianu interpolacyjnego jako kombinację liniową wielomianów Lagrange'a i wartości funkcji interpolowanej w węzłach. W ten sposób uzyskuje się wynik w postaci wielomianu

$$\begin{aligned} W(x) = l_0y_0 + l_1y_1 + l_2y_2 + l_3y_3 + l_4y_4 = y_0\frac{x\ (x-1)\ (x+1)\ (x-2)}{24} + \\ + y_1\left(-\frac{x\ (x-1)\ (x-2)\ (x+2)}{6}\right) + y_2\frac{(x-1)\ (x+1)\ (x-2)\ (x+2)}{4} + \\ + y_3\left(-\frac{x\ (x+1)\ (x-2)\ (x+2)}{6}\right) + y_4\frac{x\ (x-1)\ (x+1)\ (x+2)}{24} = \frac{x^3}{5} + 1 \end{aligned} \quad (2.14)$$

Rysunek 2.3 przedstawia kształt poszczególnych wielomianów Lagrange'a oraz wykres finalnego wielomianu interpolacyjnego. Widoczna jest znana właściwość wielomianów Lagrange'a, iż wielomian l_i przyjmuje wartość jeden dla węzła x_i, zaś dla pozostałych węzłów zeruje się – co zostało zaznaczone na poszczególnych wykresach.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 2.3. Zasada interpolacji Lagrange'a. Krzywe $y = l_i(x)$, $i = 0, 1, 2, 3, 4$ są wielomianami Lagrange'a

Jak widać na powyższym przykładzie, proces budowy wielomianów Lagrange'a l_i jest dosyć prosty. Natomiast końcowe obliczenia wymagają dużej uwagi i skupienia, ze względu na dużą łatwość popełnienia błędu rachunkowego.

2.2. Podsumowanie

W rozdziale przedstawiono zarys problematyki interpolacji. Te podstawowe informacje pozwalają na samodzielne rozwiązywanie prostych problemów spotykanych w praktyce inżynierskiej. Warto uwypuklić zalety poszczególnych metod interpolacji. Metoda Newtona jest bardzo przydatna do obliczeń prowadzonych na papierze. Stosunkowo łatwo się ją implementuje numerycznie. Gdy zachodzi konieczność dodania węzła, co może mieć miejsce przy obliczeniach z adaptacyjną zmianą długości kroku, nie trzeba na nowo wyliczać całej tablicy ilorazów różnicowych, a jedynie doliczyć ostatnie wartości w wierszach. Natomiast metoda Lagrange'a bardzo dobrze się wpisuje w rozważania teoretyczne, gdy konieczne jest zbudowanie wielomianu na podstawie znajomości położenia węzłów. W ten sposób można wyprowadzać wzory, między innymi, na całkowanie i różniczkowanie numeryczne. W obliczeniach prowadzonych na papierze zazwyczaj mniej pracochłonna jest metoda Newtona.

Przedstawione zostały najprostsze i najczęściej stosowane metody interpolacji. Wśród pominiętych, chociaż bardzo interesujących z praktycznego punktu widzenia, zagadnień należy wspomnieć interpolację funkcjami sklejanymi, interpolację Hermite'a czy interpolację trygonometryczną. Czytelnik zainteresowany tematem może znaleźć o wiele więcej informacji na temat interpolacji w literaturze przedmiotu [1, 4, 5, 13, 15, 23, 33].

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 3

Rozwiązywanie równań nieliniowych

3.1. Wstęp

Jednym z częściej występujących zagadnień praktycznych jest rozwiązywanie równań. Szczególnie równania nieliniowe mogą sprawiać różnego rodzaju problemy przy rozwiązywaniu. Gdy nie udawało się znaleźć analitycznego rozwiązania, podejmowano próby znalezienia rozwiązań przybliżonych. Tak powstawały metody iteracyjne polegające na znajdowaniu coraz lepszych przybliżeń rozwiązań. Ponieważ sam problem rozwiązywania równań nieliniowych jest dosyć stary, również metody rozwiązywania go mają swój wiek. Stosowana obecnie metoda Newtona pochodzi z końca XVII wieku. Następne lata przyniosły dalszy postęp w konstruowaniu metod iteracyjnych, i tak już 1865 roku Ernst Schröder konstruował metody iteracyjne o wykładniku zbieżności wyższym niż dwa [5].

Ponieważ metody rozwiązywania równań nieliniowych są jednymi z podstawowych narzędzi obliczeniowych, dokonany zostanie wybiórczy przegląd tych metod. Omówione zostaną najprostsze metody rozwiązywania równań nieliniowych. Przedstawiany materiał będzie dotyczył równań z jedną niewiadomą, choć nic nie stoi na przeszkodzie, by podane zależności uogólnić na większą ich liczbę. Chociaż w takim przypadku pojawiają się dodatkowe problemy związane z wyznaczaniem pochodnych i zbieżnością rozwiązania. Owe problemy nie będą głębiej analizowane w skrypcie. Zainteresowani powinni zapoznać się z pozycjami [5, 23, 30, 44].

3.2. Metoda iteracji prostej

Metoda iteracji prostej jest najprostszym przedstawicielem szerokiej klasy metod iteracyjnych. Bywa także zwana metodą kolejnych przybliżeń lub też iteracją Banacha. Z uwagi na słabą zbieżność, w podstawowej postaci nie jest zbyt często stosowana. Natomiast z dydaktycznego punktu widzenia jest bardzo przydatna do demonstracji pewnych istotnych cech metod iteracyjnych oraz jest najogólniejszą metodą iteracyjną rozwiązywania problemów nieliniowych.

Sformułowanie problemu. Rozwiązać równanie nieliniowe w postaci

$$f(x) = 0 \tag{3.1}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązanie. Aby rozwiązać tak postawiony problem, należy równanie przekształcić tak, by z lewej strony znaku równości występowała niewiadoma

$$x = g(x) \tag{3.2}$$

gdzie $g(x)$ jest funkcją ciągłą wynikającą z przekształcenia równania (3.1). Wybiera się wartość początkową x_0, a następnie wyznacza kolejne przybliżenia rozwiązania jako wartości funkcji $g(x)$ w następujący sposób

$$\begin{aligned} x_1 &= g(x_0) \\ x_2 &= g(x_1) \\ &\vdots \quad\quad \vdots \\ x_i &= g(x_{i-1}) \end{aligned} \tag{3.3}$$

Jeżeli ciąg wartości $x_0, x_1, \ldots, x_n$ dąży do granicy x^*, to $g(x_{i-1})$ dąży do granicy $g(x^*)$. Zatem w granicy uzyska się wynik

$$x^* = g(x^*) \tag{3.4}$$

czyli x^* jest rozwiązaniem (pierwiastkiem) równania (3.2), a jednocześnie równania (3.1). Zasadniczym problemem staje się tym samym kwestia zbieżności ciągu wartości kolejnych przybliżeń rozwiązania x_i. Niech funkcja $g(x)$ ma ograniczoną przez liczbę M pochodną na przedziale $E = \langle x_i, x_j \rangle$. Z twierdzenia o wartości średniej wynika, że

$$g(x_i) - g(x_j) = (x_i - x_j)g'(\xi), \qquad \xi \in E \tag{3.5}$$

z czego z kolei wynika

$$|g(x_i) - g(x_j)| \leqslant M\,|x_i - x_j| \tag{3.6}$$

Na tej podstawie można sformułować twierdzenie o zbieżności metody iteracji prostej w następującej postaci [30]:

Twierdzenie. *Jeżeli $g(x)$ jest funkcją różniczkowalną i ma w otoczeniu pierwiastka x^* pochodną ograniczoną przez liczbę $M < 1$, to ciąg $\{x_i\}$ jest zbieżny do granicy x^*.*

Ponieważ w praktycznych przypadkach, często okazuje się, że metoda iteracji prostej nie jest zbieżna lub też jej zbieżność jest zbyt wolna, należy szukać innych sposobów rozwiązywania tychże problemów. Gdy zatem metoda iteracji prostej nie jest zbieżna, można skorzystać z jej modyfikacji pozwalającej ulepszyć zbieżność, zwanej nadrelaksacją, która zostanie omówiona w następnym podrozdziale.

Metoda iteracji prostej wynika wprost z twierdzenia Banacha o kontrakcji. Rozwiązanie równania $x = g(x)$ jest punktem stałym odwzorowania $g(x)$.

Błąd przybliżenia. W celu wyznaczenia błędu metody, z ciągu kolejnych rozwiązań skonstruowany zostanie szereg

$$\begin{cases} u_0 = x_0 \\ u_n = x_n - x_{n-1}, \quad \text{dla } n = 1, 2, 3, \ldots \end{cases} \tag{3.7}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tak określony szereg ma interesujące właściwości:

— x_n jest sumą częściową pierwszych $n+1$ wyrazów szeregu, czyli $u_0+u_1+\ldots+u_n = x_n$, z czego wynika, że jeżeli szereg i ciąg są zbieżne, to mają tą samą granicę x^*,
— $x^* - x_n = u_{n+1} + u_{n+2} + \ldots$, czyli różnica $x^* - x_n$ jest resztą szeregu $\sum u_n$.

Zgodnie z drugą właściwością szeregu $\sum u_n$ błąd popełniany przy rozwiązaniu metodą iteracji prostej jest równy reszcie szeregu. Ponieważ zgodnie z zależnością (3.6) $|u_{n+p}| < M^p|u_n|$, można napisać

$$|x^* - x_n| \leqslant u_n[M + M^2 + \ldots] \tag{3.8}$$

zatem

$$|x^* - x_n| \leqslant \frac{M}{1-M}|u_n| \tag{3.9}$$

Aby wyznaczyć wartość M, można wykorzystać zależność (3.6)

$$M \approx \left|\frac{g(x_{n-1}) - g(x_{n-2})}{x_{n-1} - x_{n-2}}\right| = \left|\frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}\right| \tag{3.10}$$

Para zależności (3.9) oraz (3.10) pozwala oszacować wielkość błędu popełnianego przy rozwiązaniu równania metodą iteracji prostej dla aktualnej iteracji. Korzystając z tej informacji, można próbować wyznaczać wartość pierwiastka z dowolną dokładnością, zależną od liczby cyfr dokładnych w reprezentacji maszynowej.

3.2.1. Interpretacja geometryczna

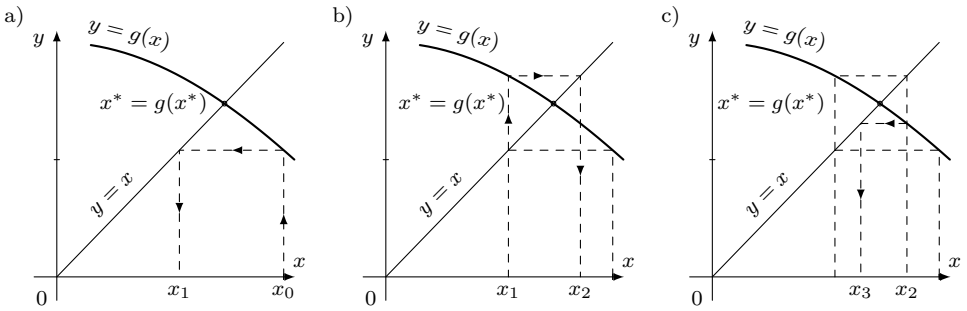
Aby graficznie rozwiązać równanie (3.2), należy narysować funkcje przedstawiające lewą i prawą stronę równania, czyli odpowiednio $y = x$ oraz $y = g(x)$. Współrzędna x punktu przecięcia będzie poszukiwanym rozwiązaniem problemu x^*. Rysunek 3.1 przedstawia poszukiwanie rozwiązania w postaci graficznej. Rysunek 3.1a pokazuje wyznaczanie wartości drugiego przybliżenia x_1 na podstawie arbitralnie przyjętej wartości x_0. Na rysunku 3.1b strzałkami zaznaczono kierunki poszukiwań kolejnego przybliżenia x_2 na podstawie uzyskanego x_1. Ostatni – rysunek 3.1c – przedstawia kolejne, tym razem czwarte przybliżenie x_3. Taki tok postępowania prowadzi się aż do momentu uzyskania wystarczającej dokładności wyznaczenia rozwiązania x^*.

Na rysunku 3.2 pokazano przykładowe rozwiązania metodą iteracji prostej. Dwa z nich (górne) przedstawiają przypadki rozbieżne, dwa dolne ukazują problemy, dla których metoda jest zbieżna. Obok odpowiednich wykresów opisano zakresy wartości pochodnej $g'(x)$ dla poszczególnych przypadków.

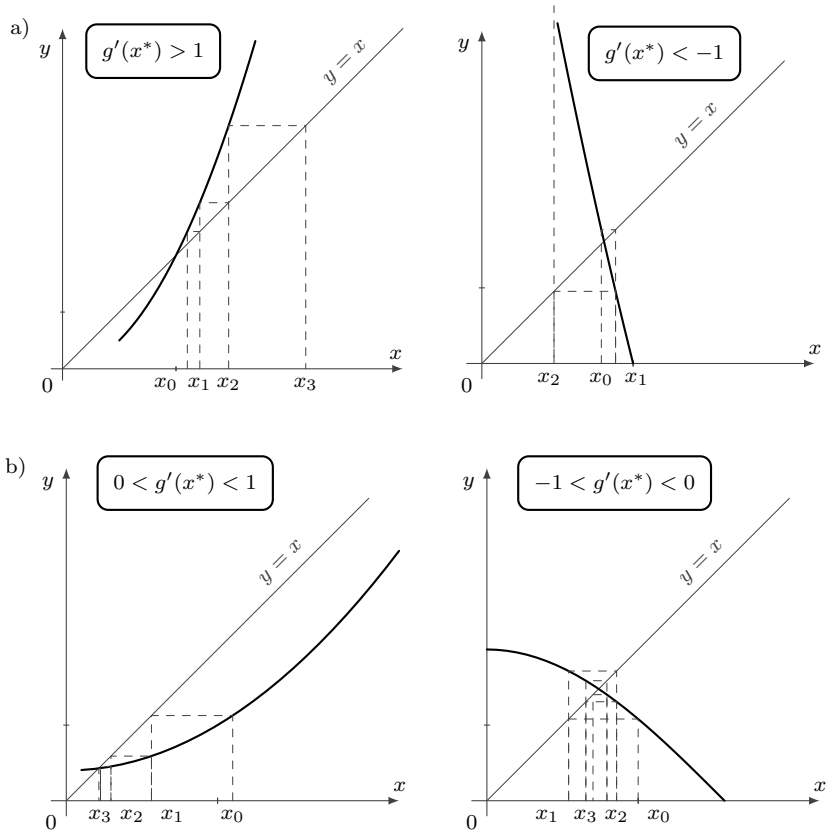
3.2.2. Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać równanie $x - \cos x = 0$ z dokładnością Δx lepszą niż 0,001, zaczynając od punktu początkowego $x_0 = 1$. Wykorzystać metodę iteracji prostej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 3.1. Interpretacja geometryczna metody iteracji prostej (opis w tekście)



Rysunek 3.2. Rozbieżne a) oraz zbieżne b) przypadki w metodzie iteracji prostej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

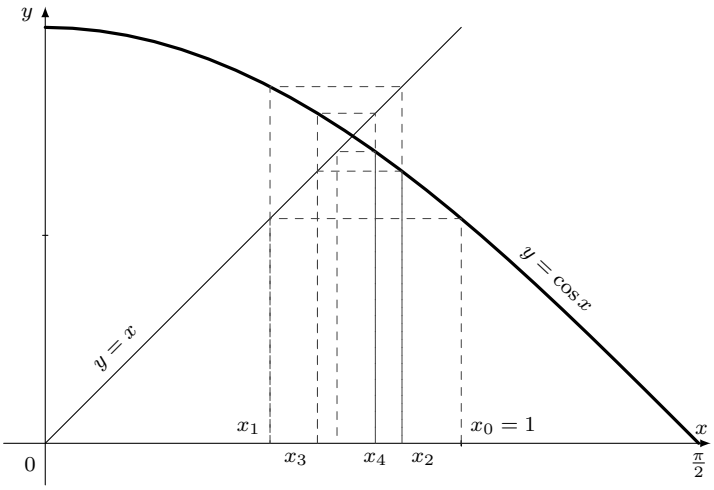
Rozwiązanie. Równanie należy przekształcić do postaci $x = g(x)$

$$x - \cos x = 0 \Leftrightarrow x = \cos x$$

Następny krok polega na zbadaniu zbieżności metody dla uzyskanego równania

$$g(x) = \cos x, \text{ co daje } g'(x) = -\sin x \Rightarrow |g'(x)| < 1$$

Ten wynik świadczy o tym, że dla tego równania metoda iteracji prostej, w obszarze rozwiązania, będzie zbieżna. To że będzie zbieżna, nie oznacza, że pozwoli efektywnie wyznaczyć wartość rozwiązania tego równania.



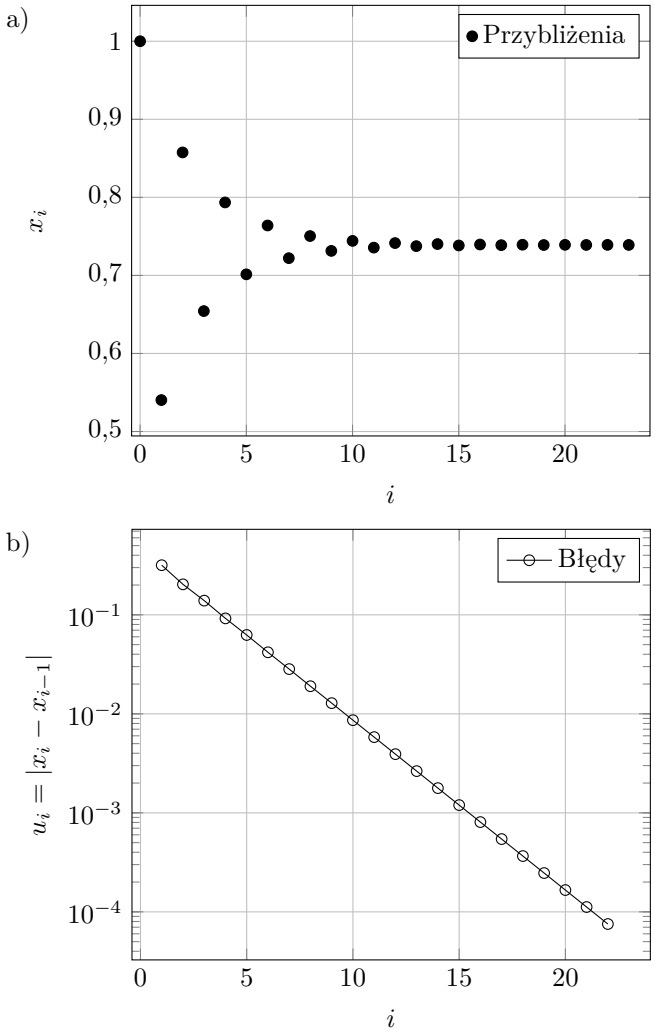
Rysunek 3.3. Rozwiązanie równania $x = \cos x$ metodą iteracji prostej – kolejne przybliżenia w otoczeniu rozwiązania

Rysunek 3.3 ilustruje rozwiązywany problem. Zaznaczono na nim kolejne przybliżenia rozwiązania równania. Przedstawione na tym rysunku kolejne przybliżenia zostały poniżej zestawione w postaci numerycznej.

$$\begin{array}{lll} x_1 = 0{,}54030 & x_7 = 0{,}72210 & x_{13} = 0{,}73751 \\ x_2 = 0{,}85755 & x_8 = 0{,}75042 & x_{14} = 0{,}74015 \\ x_3 = 0{,}65429 & x_9 = 0{,}73140 & x_{15} = 0{,}73837 \\ x_4 = 0{,}79348 & x_{10} = 0{,}74424 & x_{16} = 0{,}73957 \\ x_5 = 0{,}70137 & x_{11} = 0{,}73560 & x_{17} = 0{,}73876 \\ x_6 = 0{,}76396 & x_{12} = 0{,}74143 & x_{18} = 0{,}73930 \end{array}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Na podstawie wyznaczonych przybliżeń rozwiązania sporządzone zostały wykresy zamieszczone na rysunku 3.4.



Rysunek 3.4. Kolejne przybliżenia rozwiązania a) i różnice między nimi b)

Rysunek 3.4a przedstawia wartości rozwiązania równania uzyskane w kolejnych iteracjach, w ten sposób łatwo można zilustrować zbieżność metody dla rozpatrywanego przykładu. Rysunek 3.4b przedstawia wartość bezwzględną różnicy między kolejnymi przybliżeniami rozwiązania, jako miarę błędu bezwzględnego. Dzięki użyciu skali logarytmicznej łatwo można zauważyć liniową zbieżność metody.

Powyższy przykład jest o tyle ciekawy, iż kolejne przybliżenia lokowane są po przeciwnych stronach rozwiązania x^*. W ten sposób można uzasadnić praktycznie

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

stosowaną metodę określenia błędu rozwiązania δ, jako różnicę między kolejnymi przybliżeniami $\delta \approx |x_i - x_{i-1}|$.

Błąd ostatecznego rozwiązania na podstawie zależności (3.9) i (3.10) wynosi

$$M \approx \left| \frac{0{,}73876 - 0{,}73957}{0{,}73957 - 0{,}73837} \right| = 0{,}675$$

$$|u_n| = |0{,}73876 - 0{,}73957| = 0{,}00081$$

$$|x^* - x_n| \leqslant \frac{0{,}675}{1 - 0{,}675} 0{,}00081 = 0{,}00168$$

Zatem rozwiązanie problemu wynosi $x \approx 0{,}73876 \pm 0{,}00168$. Z ilustracji geometrycznej wynika, iż przedstawione oszacowanie błędu jest nadmiarowe. Rozmieszczenie kolejnych przybliżeń wskazuje na znacznie mniejszy błąd rozwiązania.

Jak widać z rysunku 3.4, metoda iteracji prostej jest zbieżna liniowo, co w praktyce oznacza, iż jest wolno zbieżna.

3.3. Nadrelaksacja – poprawianie zbieżności

Sformułowanie problemu. W otoczeniu poszukiwanego pierwiastka x^* funkcja $g(x)$ nie spełnia warunku zbieżności, czyli pochodna $|g'(x)| > 1$. Należy tak zmodyfikować funkcję $g(x)$, by realizując podobny schemat obliczeniowy, otrzymać pierwiastek o tej samej wartości x^*, a nowa funkcja $g_1(x)$ gwarantowała zbieżność. Wychodząc ze znanej zależności

$$x = g(x)$$

i dodając obustronnie do powyższego równania αx (przy założeniu $\alpha \neq -1$), otrzymuje się

$$\alpha x + x = \alpha x + g(x)$$

$$x(1 + \alpha) = \alpha x + g(x)$$

czyli można napisać (przy powyższym założeniu)

$$x = \frac{\alpha x + g(x)}{1 + \alpha} = g_1(x)$$

Dla porządku należy wprowadzić indeksy, by zaakcentować iteracyjną postać powyższej zależności

$$x_{i+1} = \frac{\alpha x_i + g(x_i)}{1 + \alpha} = g_1(x_i) \tag{3.11}$$

Ostatecznie schemat iteracyjny wygląda następująco

$$\begin{aligned} x_0 &- \text{przyjęte arbitralnie} \\ x_1 &= g_1(x_0) \end{aligned}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\begin{aligned} x_2 &= g_1(x_1) \\ \vdots & \quad \vdots \\ x_i &= g_1(x_{i-1}) \end{aligned}$$

Kryterium zbieżności pozostaje identyczne z tym dla metody iteracji prostej, przy czym dotyczy funkcji $g_1(x)$, czyli $|g_1'(x)| < 1$, gdzie

$$g_1'(x) = \frac{\alpha + g'(x)}{1 + \alpha} \tag{3.12}$$

Jak widać z równania (3.12), przez zmianę wartości parametru α można zmieniać wartość pochodnej $g_1'(x)$ aż do spełnienia warunku zbieżności. Warto zauważyć, że metoda iteracji prostej jest najszybciej zbieżna w sytuacji, gdy $g(x)$ jest funkcją stałą, czyli $g'(x) = 0$ w pobliżu pierwiastka x^*. Zatem optymalną, pod względem szybkości zbieżności wartość parametru α można określić jako

$$\alpha = -g'(x^*),$$

przy czym można wziąć wartość zbliżoną do pierwiastka, nie tracąc wiele na szybkości zbieżności metody.

Warto zapamiętać następujące wnioski dotyczące metody nadrelaksacji:

— gdy metoda iteracji prostej jest zbieżna liniowo do pojedynczego rozwiązania, nadrelaksacja zbiega się do tego samego rozwiązania kwadratowo,
— jeżeli rząd zbieżności metody iteracji prostej wynosi $p \geqslant 2$ i jest zbieżny do pojedynczego rozwiązania, to nadrelaksacja zbiega się do tego samego rozwiązania i ma rząd zbieżności $2p - 1$.

3.3.1. Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać równanie $x - \cos x = 0$ z dokładnością Δx lepszą niż 0,001, zaczynając od punktu początkowego $x_0 = 1$. Wykorzystać metodę iteracji prostej z nadrelaksacją.

Rozwiązanie. Ilustrację problemu przedstawia rysunek 3.3. Należy przekształcić równanie do postaci $x = g_1(x)$, gdzie

$$g_1(x) = \frac{\alpha x + g(x)}{1 + \alpha}$$

$$x - \cos x = 0 \Leftrightarrow x = \cos x \Leftrightarrow g(x) = \cos x$$

oraz optymalna wartość parametru $\alpha = -g'(x^*)$. Z rysunku 3.3 można odczytać przybliżoną wartość rozwiązania jako $x^* \approx 0{,}75$, czyli $\alpha \approx 0{,}682$

$$x_{i+1} = \frac{\alpha x_i + g(x_i)}{1 + \alpha} = \frac{0{,}682 x_i + \cos x_i}{1{,}682}$$

Wartości kolejnych przybliżeń oraz oszacowania błędów rozwiązania z użyciem nadrelaksacji zestawiono poniżej.

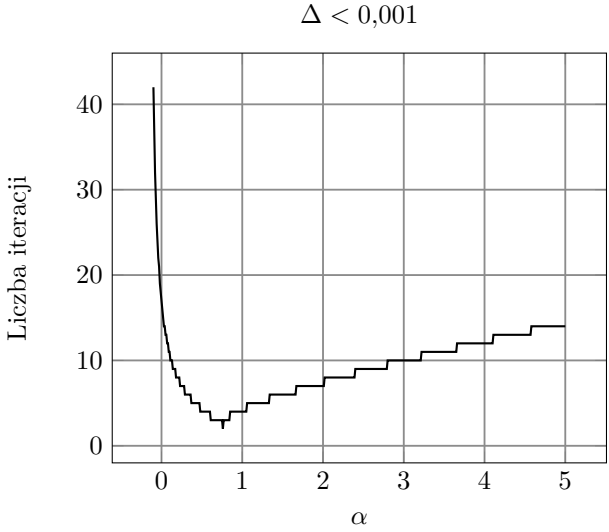
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

x_i	$\lvert x_{i-1} - x_i \rvert$
$x_0 = 1,0$	
$x_1 = 0{,}72670$	$0{,}2733$
$x_2 = 0{,}73899$	$1{,}2294 \cdot 10^{-2}$
$x_3 = 0{,}73908$	$9{,}52 \cdot 10^{-5}$

Jak widać, rozwiązanie z wymaganą dokładnością wynosi $x = 0{,}739$. Powyższy przykład dobrze ilustruje poprawienie szybkości zbieżności metody iteracji prostej. Metodą iteracji prostej rozwiązanie równania z zadaną dokładnością uzyskano w siedemnastej iteracji, natomiast po zastosowaniu nadrelaksacji w trzeciej iteracji.

3.3.2. Przykład obliczeniowy

Sformułowanie problemu. Zbadać zachowanie metody iteracji prostej z nadrelaksacją przy rozwiązywaniu równania $x - \cos x = 0$ z dokładnością Δx lepszą niż 0,001, przyjmując $x_0 = 1$ oraz parametr α z przedziału $\langle -0{,}1; 5 \rangle$.



Rysunek 3.5. Zależność liczby iteracji od parametru α dla zadanej dokładności Δ

Rozwiązanie. Podobnie jak w poprzednim przykładzie, ilustrację problemu przedstawia rysunek 3.3. Po przeprowadzeniu obliczeń, zależność liczby iteracji niezbędnych do osiągnięcia wymaganej dokładności od parametru α można zilustrować rysunkiem 3.5.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

3.4. Metoda Δ^2 Aitkena

Bardzo ciekawy sposób przyspieszania zbieżności ciągu o zbieżności liniowej opracował Alexander Aitken w roku 1926.

Jeżeli założy się, że a_n jest ciągiem wartości uzyskiwanych z metody numerycznej i jest zbieżny liniowo, to można zbudować ciąg b_n zbieżny kwadratowo do tej samej wartości o następującej konstrukcji

$$b_n = a_n - \frac{(\Delta a_n)^2}{\Delta^2 a_n} \tag{3.13}$$

gdzie $\Delta a_n = a_{n+1} - a_n$, zaś $\Delta^2 a_n = \Delta a_{n+1} - \Delta a_n = a_{n+2} - 2a_{n+1} + a_n$.

Taki sposób przyspieszania zbieżności ciągu można zastosować na przykład do wyznaczania wartości własnych czy rozwiązywania równań nieliniowych. Metoda Aitkena bardzo dobrze sprawdza się przy przyspieszaniu zbieżności metody iteracji prostej. Poniżej zostanie przedstawiony sposób postępowania w takim przypadku. Jeżeli metoda iteracji prostej jest zbieżna liniowo do rozwiązania x^*, to można zapisać

$$g(x_i) - x^* = \lambda(x_i - x^*)$$

co można przekształcić do postaci

$$x^* = \frac{g(x_i) - \lambda x_i}{1 - \lambda}$$

po dodaniu i odjęciu x_i od licznika można uzyskać następującą postać

$$x^* = \frac{g(x_i) - \lambda x_i u + x_i - x_i}{1 - \lambda} = \frac{x_i(1 - \lambda)}{1 - \lambda} + \frac{g(x_i) - x_i}{1 - \lambda}$$

czyli

$$x^* = x_i + \frac{g(x_i) - x_i}{1 - \lambda} = x_i + \frac{x_{i+1} - x_i}{1 - \lambda}$$

Aby wyznaczyć nieznaną wartość λ, można wykorzystać liniową zbieżność ciągu x_i i napisać dla dwóch kolejnych przybliżeń

$$g(x_{i+1}) - x_{i+1} = \lambda(x_{i+1} - x_i)$$

czyli

$$\lambda = \frac{g(x_{i+1}) - x_{i+1}}{x_{i+1} - x_i}$$

Zatem zależność iteracyjna pozwalająca rozwiązać równanie nieliniowe przyjmuje postać

$$x^* = x_i + \frac{x_{i+1} - x_i}{1 - \frac{g(x_{i+1}) - x_{i+1}}{x_{i+1} - x_i}} = x_i - \frac{(x_{i+1} - x_i)^2}{g(x_{i+1}) - 2x_{i+1} + x_i}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Aby uzyskać końcową postać metody Aitkena dla metody iteracji prostej, należy dokonać dwóch podstawień. Pierwsze podstawienie to $x_{i+1} = g(x_i)$, natomiast drugie to $g(x_{i+1}) = g(g(x_i))$. Końcowa postać zależności wygląda następująco

$$x_{i+1} = x_i - \frac{(g(x_i) - x_i)^2}{g(g(x_i)) - 2g(x_i) + x_i} \tag{3.14}$$

Opisana metoda Aitkena znacznie poprawia szybkość zbieżności metod iteracyjnych. Jeżeli metoda iteracyjna jest zbieżna liniowo do pojedynczego zera, metoda Aitkena jest zbieżna kwadratowo. Jak już wspomniano, można ją stosować nie tylko do rozwiązywania równań nieliniowych.

Metodę podobną do metody Aitkena można znaleźć w pracach Seki Takakazu, japońskiego matematyka z drugiej połowy XVII wieku. Wykorzystywał tę metodę do obliczenia wartości liczby π z dokładnością do 10. miejsca po przecinku.

3.5. Związek metody iteracji prostej z metodą Newtona

Aby rozwiązać równanie $f(x) = 0$ metodą iteracji prostej, można dodać stronami składnik zawierający x, na przykład αx. Dla $\alpha \neq 0$ można zatem napisać

$$f(x) = 0 \quad \Leftrightarrow \quad f(x) + \alpha x = \alpha x$$

czyli

$$x = g_2(x), \quad \text{gdzie } g_2(x) = x + \frac{f(x)}{\alpha}$$

Takie równanie można rozwiązać metodą iteracji prostej, zapisując je w postaci rekurencyjnej

$$x_{i+1} = g_2(x_i) \tag{3.15}$$

Ze względu na szybkość zbieżności korzystne jest to, by $|g_2'(x)|$ było bliskie zeru; można zatem zapisać

$$g_2'(x) = 1 + \frac{f'(x)}{\alpha} = 0$$

wobec tego dla $x = x^*$, $\alpha = -f'(x^*)$. Ponieważ wartość pochodnej $f'(x^*)$ zazwyczaj nie jest znana, wygodnie jest przyjąć wartość pochodnej w aktualnie wyznaczonym przybliżeniu rozwiązania, czyli $\alpha = -f'(x_i)$. Po wstawieniu tak wyznaczonej wartości α do równania (3.15) otrzyma się zależność

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

która stanowi formułę metody Newtona. Identyczną formułę można wyprowadzić w inny sposób, co zostanie pokazane w następnym podrozdziale.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

3.6. Metoda Newtona

Bardzo wydajną metodą rozwiązywania równań nieliniowych jest metoda Newtona, zwana też metodą linearyzacji czy stycznych. Obecnie jest ona z powodzeniem używana do rozwiązywania nieliniowych problemów inżynierskich i w różnych odmianach można ją znaleźć w każdym chyba systemie do obliczeń inżynierskich. Jest ceniona ze względu na swą szybką zbieżność.

Sformułowanie problemu. Rozwiązać równanie nieliniowe $f(x) = 0$. Rozważa się pierwiastek jednokrotny. Druga pochodna $f''(x)$ funkcji $f(x)$ jest ciągła w rozpatrywanym przedziale. Przyjmuje się przybliżoną wartość rozwiązania x_i. Niech miejsce zerowe będzie odległe o h od przybliżenia $x^* = x_i + h$. Aby rozwiązać to równanie, funkcję $f(x)$ w pobliżu przybliżenia rozwiązania x_i rozwija się w szereg Taylora

$$f(x_i + h) = f(x_i) + h\frac{f'(x_i)}{1!} + \frac{h^2}{2!}f''(\xi) \tag{3.16}$$

gdzie ξ jest wartością z przedziału $\langle x_i, x^* \rangle$. Skoro $x^* = x_i + h$, to lewa strona powyższej równości jest równa zero. Po ograniczeniu się do dwóch pierwszych wyrazów rozwinięcia w szereg można uzyskać

$$0 = f(x_i) + hf'(x_i) \Leftrightarrow h = -\frac{f(x_i)}{f'(x_i)} \tag{3.17}$$

Powyższe wyrażenie pozwala na wyznaczenie poprawki h, po podstawieniu tej poprawki do wzoru $x^* = x_i + h$ można napisać

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{3.18}$$

Zależność (3.18) opisuje metodę Newtona służącą do rozwiązywania równań nieliniowych. Istotne jest założenie dotyczące jednokrotności poszukiwanego pierwiastka, ponieważ przy pierwiastkach wielokrotnych zeruje się sama funkcja i jej pierwsza pochodna.

3.6.1. Interpretacja geometryczna

Na rysunku 3.6 przedstawiono interpretację geometryczną metody Newtona. Niech będzie poszukiwane miejsce zerowe x^*. Poszukiwania rozpoczyna się z punktu x_0. Prowadzi się styczną do wykresu funkcji $f(x)$ w punkcie $(x_0, f(x_0))$. Na przecięciu z osią OX znajduje się kolejne (lepsze) przybliżenie miejsca zerowego x_1.

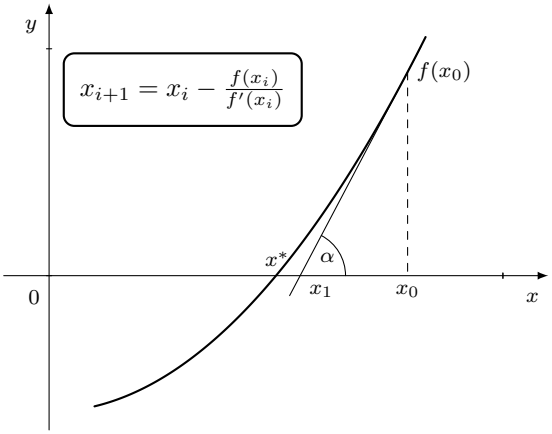
Aby wyznaczyć wartość przybliżenia x_1, należy skorzystać ze znanej informacji, że wartość pochodnej funkcji w punkcie $f'(x_0)$ jest równa współczynnikowi kierunkowemu stycznej do funkcji w tym punkcie. Współczynnik kierunkowy prostej to nic innego jak tangens kąta nachylenia stycznej do osi OX, czyli można napisać

$$f'(x_0) = \operatorname{tg} \alpha$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Algorytm 3.1. Metoda Newtona

Require: x_0, tol, $f'(x)$
  repeat
    $a \leftarrow f(x_0)$
    $b \leftarrow f'(x_0)$
    if $b \neq 0$ **then**
      $h \leftarrow a/b$
    else
      return -1
    end if
    $x_0 \leftarrow x_0 - h$
  until $abs(h) \geqslant tol$
  return x_0



Rysunek 3.6. Interpretacja geometryczna metody Newtona

W trójkącie prostokątnym $x_1x_0f(x_0)$ tangens kąta α wynosi

$$\operatorname{tg}\alpha = \frac{f(x_0)}{x_0 - x_1}$$

Można zatem zapisać następujące przekształcenie

$$x_0 - x_1 = \frac{f(x_0)}{\operatorname{tg}\alpha} \Leftrightarrow x_0 - x_1 = \frac{f(x_0)}{f'(x_0)} \Leftrightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

co wprost prowadzi do znanej zależności opisującej metodę Newtona

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Przedstawioną procedurę iteracyjną powtarza się do uzyskania zadowalającej dokładności mierzonej długością kroku metody, czyli ilorazu $f(x_0)/f'(x_0)$. W większości

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

problemów takie kryterium zakończenia obliczeń jest wystarczające, niemniej gdy pierwsza pochodna $f'(x)$ w okolicach rozwiązania przyjmuje niewielkie wartości, może być mylące. Nieco więcej na temat warunków skończenia obliczeń powiedziano w dalszej części niniejszego rozdziału, w podrozdziale 3.9. Szerzej to zagadnienie jest opisane na przykład w pozycji [5].

3.6.2. Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać równanie $x - \cos x = 0$ z dokładnością Δx lepszą niż 0,001, zaczynając od punktu początkowego $x_0 = 1$. Wykorzystać metodę Newtona.

Rozwiązanie. Zgodnie z treścią zadania $\Delta x = 0{,}001$, $x_0 = 1$, $f(x) = x - \cos x$, $f'(x) = 1 + \sin x$. Wobec tego metoda Newtona dla tego równania może być zapisana w postaci następującej zależności rekurencyjnej

$$x_{i+1} = x_i - \frac{x_i - \cos x_i}{1 + \sin x_i}$$

Kolejne przybliżenia rozwiązania przedstawiają się, jak przedstawiono w tablicy poniżej.

x_i	$\lvert x_i - x_{i+1}\rvert$
$x_0 = 1{,}0$	
$x_1 = 0{,}75036$	$0{,}24964$
$x_2 = 0{,}73911$	$1{,}125 \cdot 10^{-2}$
$x_3 = 0{,}73909$	$2{,}77 \cdot 10^{-5}$

Podobnie jak przy wykorzystaniu nadrelaksacji wystarczyły trzy iteracje do uzyskania rozwiązania z wymaganą dokładnością. Na podstawie wyników obliczeń można zauważyć szybką zbieżność. W kolejnych iteracjach następuje szybkie zwiększanie liczby cyfr dokładnych.

3.6.3. Zbieżność metody

Szybkość zbieżności metod iteracyjnych charakteryzuje tak zwany wykładnik zbieżności, który definiowany jest za pomocą wyrażenia

$$\lim_{n\to\infty} \frac{|(\Delta x)_{n+1}|}{|(\Delta x)_n|^p} = C$$

gdzie C jest stałą asymptotyczną błędu, $(\Delta x)_n = x^* - x_n$ jest błędem bezwzględnym rozwiązania w kroku n, zaś p jest wykładnikiem zbieżności. W zależności od

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wartości p, mówi się o różnych rodzajach zbieżności. I tak, dla $p = 1, 2, 3$ mówi się odpowiednio o zbieżności liniowej, kwadratowej i sześciennej. Dla wartości $p \geqslant 3$ mówi się o zbieżności rzędu p. Ważnym rodzajem zbieżności jest zbieżność logarytmiczna, która występuje, gdy

$$\lim_{n\to\infty} \frac{|(\Delta x)_{n+1}|}{|(\Delta x)_n|} = 1$$

czyli $p = 1$ oraz $C = 1$. W praktyce numerycznej metody, których wyniki są zbieżne liniowo lub logarytmicznie, są rzadko stosowane ze względu na trudności w uzyskaniu wymaganej dokładności rozsądnym kosztem.

Do badania szybkości zbieżności metody Newtona przyjmuje się następujące założenia: funkcja $f(x)$ ma drugą pochodną ciągłą, a poszukiwane miejsce zerowe x^* jest pojedyncze. Niech $(\Delta x)_n$ będzie błędem przybliżenia rozwiązania

$$(\Delta x)_n = x^* - x_n$$

Rozwijając funkcję $f(x)$ w szereg Taylora w otoczeniu przybliżenia pierwiastka x_n, można uzyskać zależność

$$f(x^*) = 0 = f(x_n) + (x^* - x_n)f'(x_n) + \frac{1}{2}(x^* - x_n)^2 f''(\xi) \tag{3.19}$$

gdzie ξ jest wartością z wnętrza przedziału $\langle x^*, x_n\rangle$. Po podzieleniu przez pochodną $f'(x_n)$ otrzymuje się

$$\frac{f(x_n)}{f'(x_n)} + x^* - x_n + \frac{1}{2}\frac{(x^* - x_n)^2 f''(\xi)}{f'(x_n)} = 0$$

czyli

$$x^* - \underbrace{\left(x_n - \frac{f(x_n)}{f'(x_n)}\right)}_{x_{n+1}} = -\frac{1}{2}\frac{(x^* - x_n)^2 f''(\xi)}{f'(x_n)}$$

Zatem podstawiając za odpowiednie wyrażenia $(\Delta x)_n$ i $(\Delta x)_{n+1}$, otrzymuje się

$$(\Delta x)_{n+1} = \frac{1}{2}(\Delta x)_n^2 \frac{f''(\xi)}{f'(x_n)} \tag{3.20}$$

Uzyskany wynik należy podzielić stronami przez $(\Delta x)_n^2$, po przejściu do granicy $x_n \to x^*$, równanie przyjmuje postać

$$\lim_{x_n\to x^*} \frac{(\Delta x)_{n+1}}{(\Delta x)_n^2} = \frac{1}{2}\frac{|f''(x^*)|}{|f'(x^*)|} \tag{3.21}$$

Wartość w liczniku ułamka po prawej stronie wynika z twierdzenia o trzech ciągach. Druga potęga w mianowniku wyrażenia pod granicą oznacza, iż metoda Newtona ma wykładnik zbieżności równy 2 lub, inaczej mówiąc, że jest zbieżna kwadratowo.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Za pomocą metody Newtona można wyprowadzić przypisywany Heronowi wzór pozwalający wyznaczać wartości pierwiastka kwadratowego liczby rzeczywistej A na podstawie równania $x^2 - A = 0$

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{A}{x_n}\right)$$

W praktyce zbieżność kwadratowa oznacza, iż począwszy od pewnego przybliżenia, przy każdej następnej iteracji podwaja się liczba cyfr dokładnych.

3.7. Metoda siecznych

Bardzo podobna do metody Newtona jest metoda siecznych, która zostanie teraz omówiona. Najprościej zależności opisujące metodę siecznych można wyprowadzić, wychodząc ze wzoru na metodę Newtona (3.18). Należy zamienić pochodną w mianowniku na iloraz różnicowy

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

co pozwala uzyskać następujący wynik

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \approx x_i - \frac{f(x_i)}{\frac{f(x_i)-f(x_{i-1})}{x_i - x_{i-1}}} \tag{3.22}$$

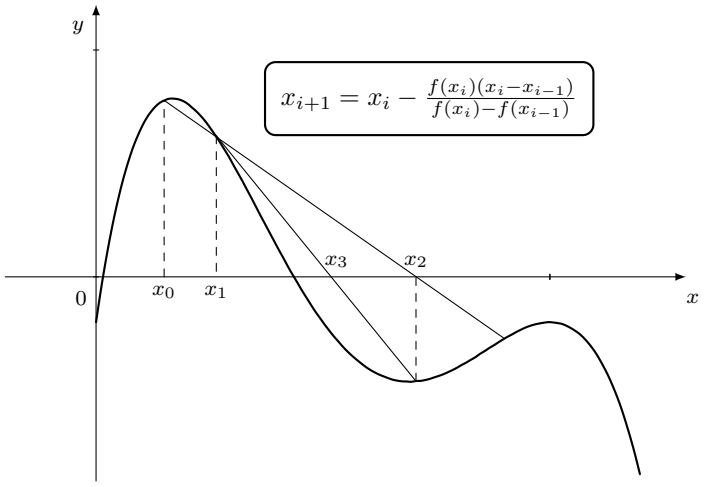
Zatem końcowa postać wzoru opisującego metodę siecznych wygląda następująco

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \tag{3.23}$$

Jak widać, aby rozpocząć obliczenia, konieczne są dwie wartości początkowe x_0 oraz x_1, zaś w każdym kroku obliczeniowym należy wyznaczyć jedną nową wartość funkcji $f(x_n)$. Strategię wyboru wartości początkowych można obrać w zasadzie dowolnie. Spotykane są, między innymi: stała różnica, stała względna różnica czy wybór losowy. W metodzie siecznych do obliczeń wybiera się zawsze dwie ostatnie wyznaczone wartości przybliżenia miejsca zerowego. Interpretacja geometryczna metody siecznych została przedstawiona na rysunku 3.7.

Metoda siecznych jest wolniej zbieżna niż metoda Newtona. Bywa używana, gdy nie zna się jawnej postaci pochodnej funkcji $f(x)$ lub też koszt obliczenia pochodnej, rozumiany jako liczba operacji arytmetycznych koniecznych do wyznaczenia wartości tej pochodnej, jest większy niż koszt obliczenia ilorazu różnicowego we wzorze (3.23). Ściśle rzecz ujmując, w każdej iteracji metody Newtona wyznacza się jedną wartość funkcji i jedną wartość pochodnej, w metodzie siecznych wyznacza się jedną wartość funkcji i iloraz różnicowy, czyli jedno dzielenie i dwa odejmowania. Jeżeli iloraz kosztu obliczenia pochodnej $f'(x)$ i kosztu obliczenia ilorazu różnicowego oznaczy się przez θ, to należy stosować metodę siecznych, gdy $\theta > 0,44$ [5]. W takiej sytuacji metodą siecznych można uzyskać wynik w krótszym czasie, pomimo jej wolniejszej zbieżności.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 3.7. Interpretacja geometryczna metody siecznych

3.7.1. Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać równanie $x - \cos x = 0$ z dokładnością Δx lepszą niż 0,001, zaczynając od punktu początkowego $x_0 = 1$. Wykorzystać metodę siecznych.

Rozwiązanie. Niech $\Delta x = 0{,}001$, $x_0 = 1$, $f(x) = x - \cos x$. Metoda siecznych wymaga dwóch punktów początkowych do rozpoczęcia obliczeń, zatem niech drugi punkt to będzie $x_1 = 1{,}1$. Metoda siecznych dla tego zadania może być zapisana następująco

$$x_{i+1} = x_i - \frac{(x_i - \cos x_i)(x_i - x_{i-1})}{x_i - \cos x_i - x_{i-1} + \cos x_{i-1}}$$

Uzyskane w trakcie obliczeń kolejne przybliżenia rozwiązania zestawiono poniżej.

x_i	$\lvert x_i - x_{i+1}\rvert$
$x_0 = 1{,}0$	
$x_1 = 1{,}1$	
$x_2 = 0{,}75379$	$0{,}34621$
$x_3 = 0{,}74004$	$1{,}375 \cdot 10^{-2}$
$x_4 = 0{,}73909$	$9{,}5 \cdot 10^{-4}$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Podobnie jak przy poprzednio omawianych metodach wystarczają trzy iteracje. Trzeba zwrócić uwagę, iż rozwiązywane równanie miało niezwykle prostą postać. Zazwyczaj obserwuje się znaczne różnice w szybkości zbieżności metod rozwiązywania równań nieliniowych.

Algorytm 3.2. Metoda siecznych

Require: x_0, x_1, tol
  $b \leftarrow f(x_0)$
  repeat
    $a \leftarrow f(x_1)$
    $h \leftarrow x_1 - x_0$
    if $a - b \neq 0$ **then**
      $x_2 \leftarrow x_1 - (ah)/(a - b)$
    else
      return -1
    end if
    $x_0 \leftarrow x_1$
    $x_1 \leftarrow x_2$
    $b \leftarrow a$
  until $abs(h) \geqslant tol$
  return x_1

3.8. Regula falsi

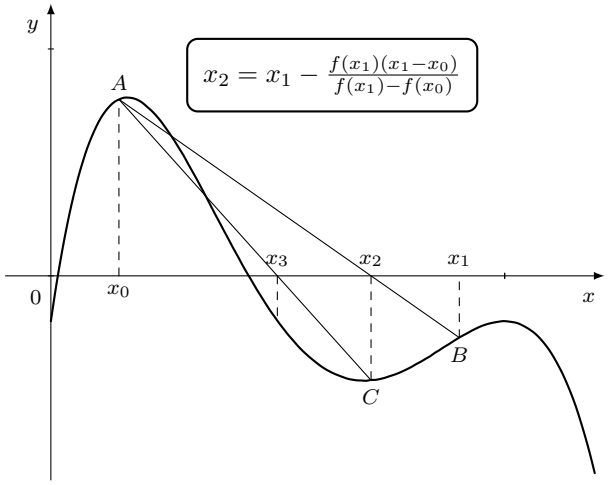
Nazwa pochodzi z łaciny i oznacza fałszywą regułę czy też fałszywą prostą. Jak będzie pokazane, sama metoda jest bardzo podobna do metody siecznych. W zasadzie jest jej niewielką modyfikacją, ale dzięki tej modyfikacji uzyskuje bardzo korzystną właściwość – jest to metoda zawsze zbieżna. Zasadniczą różnicą między nimi jest sposób wyboru wartości służących do wyznaczania kolejnego przybliżenia miejsca zerowego. W metodzie siecznych wybiera się zawsze dwa ostatnie punkty, zaś w aktualnie omawianej metodzie wybierane są zawsze punkt ostatni oraz ostatni z poprzednich, leżący po przeciwnej stronie osi OX, co oznacza, że jedna wartość jest ujemna, a druga dodatnia. Obliczenia rozpoczynają się od wyboru dwóch punktów początkowych x_0 i x_1. Ponieważ konieczne jest istnienie miejsca zerowego w przedziale $\langle x_0, x_1 \rangle$, wybrane punkty muszą spełniać warunek

$$f(x_0)f(x_1) < 0$$

Następnie, jak pokazano na rysunku 3.8, prowadzi się prostą łączącą punkty A oraz B. Przecięcie uzyskanej prostej z osią OX daje punkt x_2, który jest kolejnym przybliżeniem miejsca zerowego. Punkt x_2 można wyznaczyć z zależności (3.24)

$$x_2 = x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)} \tag{3.24}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 3.8. Interpretacja geometryczna metody regula falsi

Do następnego kroku wybiera się punkty A oraz C, w wyniku uzyskując punkt x_3. Obliczenia się powtarza aż do uzyskania wystarczającej dokładności.

Algorytm 3.3. Regula falsi

Require: x_0, x_1, tol
Ensure: $f(x_0)f(x_1) < 0$
  $b \leftarrow f(x_0)$
  repeat
    $a \leftarrow f(x_1)$
    $h \leftarrow x_1 - x_0$
    $x_2 \leftarrow (x_0 a - x_1 b)/(a - b)$
    if $f(x_0)f(x_2) < 0$ **then**
      $x_1 = x_2$
    else
      $x_0 = x_2$
    end if
  until $abs(h) \geqslant tol$
  return x_2

W pobliżu rozwiązania zbieżność tej metody jest zdecydowanie gorsza od zbieżności metody Newtona oraz nieco gorsza od zbieżności metody siecznych. Jej zasadniczą zaletą jest to, że jest zawsze zbieżna. Ze względu na słabą zbieżność nie należy jej używać w sąsiedztwie rozwiązania, a korzystać z innych metod. Za to świetnie nadaje się do wyznaczenia przybliżonej wartości pierwiastka, by następnie zwiększyć dokładność rozwiązania inną metodą, która mogłaby nie być zbieżna w większej odległości od pierwiastka.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Najstarszą istniejącą wzmiankę o metodzie regula falsi można znaleźć w indyjskim tekście Vaishali Ganit, datowanym na III wiek p.n.e. Podobnie chiński traktat *Dziewięć rozdziałów o sztukach matematycznych* wspomina o tym algorytmie, opisując jego zastosowanie do równań liniowych, co daje rozwiązanie w pierwszym kroku. Opublikowane w 1202 roku dzieło *Liber Abaci* (tłumaczone jako *Księga rachunków*) autorstwa Leonardo z Pizzy (1175–1250 r. n.e.), znanego też jako Fibonacci, wspomina o tej metodzie z powołaniem na źródła arabskie.

3.9. Warunki zakończenia obliczeń

Zakończenie obliczeń przy rozwiązywaniu równań nieliniowych może czasem nastręczać problemów. Osoby stykające się po raz pierwszy z zagadnieniami numerycznymi czasem mówią, że obliczenia należy zakończyć wtedy, gdy wartość funkcji wyniesie zero. Niestety, ze względu na zawsze obecne błędy zaokrągleń, do takiej sytuacji dochodzi bardzo rzadko. Dlatego zazwyczaj jako warunek zakończenia podaje się dopuszczalną wielkość błędu, z jakim jest wyznaczone rozwiązanie.

Poniżej zostały przedstawione wybrane z miar, które mogą być pomocne przy podejmowaniu decyzji o zakończeniu obliczeń. Jak już wspomniano, jako jedną z miar dokładności rozwiązania można przyjąć

$$|x_i - x_{i-1}| \leqslant \varepsilon$$

Inną możliwością jest przyjęcie miary względnej, w postaci

$$\left| \frac{x_i - x_{i-1}}{x_i} \right| \leqslant \varepsilon_w$$

Można też badać wielkość residuum w postaci

$$\left| \frac{f(x_i)}{f(x_0)} \right| \leqslant \varepsilon_r$$

Wielkość ε występująca w powyższych zależnościach ma interpretację dokładności obliczeń bezwzględnej lub względnej. W przypadkach jednowymiarowych, gdy rozwiązuje się problemy z jedną zmienną niezależną, powyższe wzory można stosować bezpośrednio. Gdy rozwiązuje się problem wielowymiarowy, to zamiast wartości bezwzględnych należy używać norm wektorów.

Oczywiście można też ograniczyć liczbę wykonywanych iteracji, co jest raczej warunkiem zabezpieczającym przed zbyt długimi obliczeniami, gdy problem nie wykazuje dostatecznie szybkiej zbieżności i nie osiąga w zadanym czasie rozwiązania.

3.10. Przykłady praktyczne

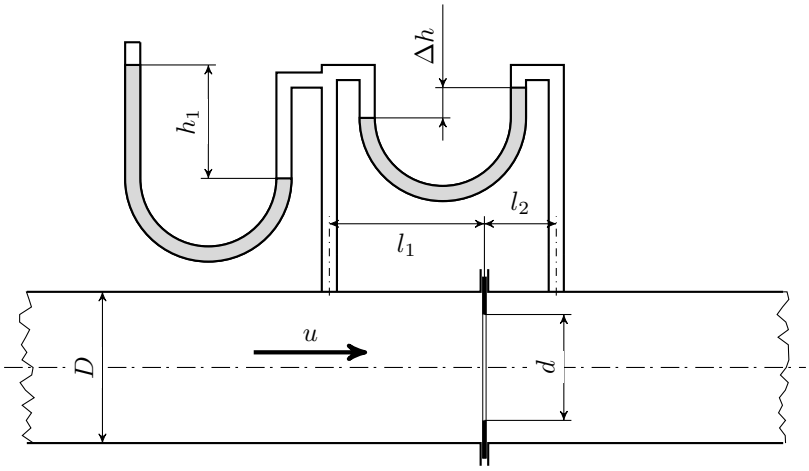
Aby przybliżyć praktyczne aspekty rozwiązywania równań nieliniowych występujących w problemach technicznych, zostanie omówionych kilka przykładów. Dla

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

ułatwienia zostaną one nieco uproszczone, zatem nie mogą stanowić wzoru postępowania dla rzeczywistych zagadnień. Mają jedynie pozwolić na demonstrację opisanych w niniejszym rozdziale metod rozwiązywania równań nieliniowych.

3.10.1. Pomiar strumienia objętości za pomocą kryzy

Wyznaczyć strumień masy powietrza przepływającego przez rurociąg. Element pomiarowy stanowi kryza. Schemat stanowiska pomiarowego przedstawiono na rysunku 3.9. Średnica rurociągu $D = 70$ mm, średnica otworu kryzy $d = 50$ mm. Nadciśnienie przed kryzą wynosi 440 Pa zaś podciśnienie za kryzą 45 Pa, wilgotność względna powietrza 42,4%, a ciśnienie atmosferyczne wynosi 98 920 Pa.



Rysunek 3.9. Schemat stanowiska pomiarowego

3.10.2. Tok obliczeń

Aby określić wielkość objętościowego natężenia przepływu za pomocą kryzy, należy wyznaczyć eksperymentalnie następujące wielkości:

— ciśnienie przed kryzą p_1,
— ciśnienie za kryzą p_2,
— temperaturę powietrza t,
— ciśnienie atmosferyczne p_b,
— wilgotność względną φ_1.

Wartości uzyskane z pomiarów oraz wymiary stanowiska wygodnie jest zebrać w tabeli. Tabela 3.1 w pierwszej części zawiera wyniki uzyskane w trakcie pomiaru. Druga część tabeli zawiera dane dotyczące wymiarów samego stanowiska.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 3.1. Dane pomiarowe do obliczeń

Wielkość	Oznaczenie	Wartość	Jednostka
Ciśnienie manometryczne przed kryzą	p_{1m}	440	Pa
Podciśnienie manometryczne za kryzą	p_{2m}	−45	Pa
Temperatura powietrza	t	25	°C
Ciśnienie barometryczne	p_b	98 920	Pa
Wilgotność względna	φ_1	48	%
Średnica rurociągu	D	70	mm
Średnica otworu kryzy	d	50	mm
Odległość do odbioru ciśn. na wlocie	l_1	70	mm
Odległość do odbioru ciśn. na wylocie	l_2	35	mm

Równanie (3.25) jest ogólnym równaniem obowiązującym dla zwężek i kryz pomiarowych i opisuje zależność strumienia masy od różnicy ciśnień.

$$q_m = \frac{C}{\sqrt{1-\beta^4}} \varepsilon \frac{\pi}{4} d^2 \sqrt{2\,\Delta p\,\rho_1} \tag{3.25}$$

Znając strumień masy, można wyznaczyć strumień objętości płynu przepływającego przez rurociąg z zależności (3.26).

$$q_V = \frac{q_m}{\rho_1} \tag{3.26}$$

gdzie C jest współczynnikiem przepływu, β to przewężenie kryzy $\beta = \frac{d}{D}$, ε jest liczbą ekspansji, Δp oznacza różnicę ciśnień przed i za kryzą, ρ_1 to gęstość powietrza, d jest średnicą otworu zwężki, zaś D to średnica rurociągu.

Z wzorów (3.25) i (3.26) wynika konieczność wyznaczenia gęstości powietrza wilgotnego. Zależy ona od temperatury otoczenia, ciśnienia atmosferycznego oraz wilgotności względnej. Można ją wyznaczyć z poniższego równania [14].

$$\rho_1 = \frac{p_1}{R\,(273{,}15 + t)} \tag{3.27}$$

gdzie R jest stałą gazową powietrza wilgotnego, której wartość można obliczyć na podstawie zależności

$$R = 462\frac{x + 0{,}622}{1 + x}, \text{ gdzie } x = 0{,}622\frac{\varphi_1 p''}{p_1 - \varphi_1 p''}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Ciśnienie pary wodnej nasyconej w temperaturze t dla zakresu $t \in (0, 30)\,^{\circ}\mathrm{C}$ można wyznaczyć z zależności

$$p'' = 3{,}47\,t^2 + 16{,}9\,t + 611{,}0$$

Następną wielkością, którą należy wyznaczyć, jest współczynnik przepływu określający zależność miedzy rzeczywistym a teoretycznym strumieniem masy lub objętości płynu. Zależność empiryczna opisująca ten współczynnik wiąże rodzaj zwężki, odległości punktów pomiaru ciśnienia przed i za zwężką oraz liczbę Reynoldsa w następujący sposób

$$\begin{aligned} C = {} & 0{,}5961 + 0{,}0261\beta^2 - 0{,}216\beta^8 + 0{,}000521\left(\frac{10^6\beta}{\mathrm{Re}_D}\right)^{0{,}7} + \\ & + (0{,}0188 + 0{,}0063\,A)\beta^{3{,}5}\left(\frac{10^6}{\mathrm{Re}_D}\right)^{0{,}3} + \left(0{,}043 + 0{,}080e^{-10L_1} - 0{,}123e^{-7L_1}\right)\cdot \\ & \cdot (1 - 0{,}11A)\cdot\frac{\beta^4}{1-\beta^4} - 0{,}031(M_2 - 0{,}8M_2^{1{,}1})\beta^{1{,}3} + 0{,}011(0{,}75 - \beta)\left(2{,}8 - \frac{D}{25{,}4}\right) \end{aligned} \tag{3.28}$$

gdzie $L_1 = l_1/D$ jest ilorazem odległości otworu impulsowego dopływowego od dopływowej powierzchni kryzy i średnicy rurociągu, $L_2 = l_2/D$ jest ilorazem odległości otworu impulsowego odpływowego od odpływowej powierzchni kryzy i średnicy rurociągu, $M_2 = \frac{2L_2}{1-\beta}$, $A = \left(\frac{19\,000\beta}{Re_D}\right)^{0{,}8}$.

Liczbę Reynoldsa oblicza się zgodnie z wzorem (3.29).

$$\mathrm{Re}_D = \frac{\rho_1 u D}{\mu_1} = \frac{\rho_1 \frac{4q_m}{\pi D^2 \rho_1} D}{\mu_1} = \frac{4q_m}{\pi\mu_1 D} \tag{3.29}$$

We wzorze (3.29) występuje wielkość μ_1, jest to lepkość dynamiczna powietrza w temperaturze t. Do jej wyznaczenia można użyć wzoru Sutherlanda.

$$\mu_1 = \mu_n \frac{1 + \frac{C_s}{273{,}15}}{1 + \frac{C_s}{T}}\sqrt{\frac{T}{273{,}15}} \tag{3.30}$$

gdzie $\mu_1 = 1{,}716 \cdot 10^{-4}$ Pa·s, μ_n jest lepkością dynamiczną powietrza w warunkach normalnych, T jest temperaturą powietrza wyrażoną w [K], C_s to stała Sutherlanda, dla powietrza wynosząca $C_s = 111$ K.

Ostatnią wielkością, którą należy wyznaczyć, by określić objętościowe natężenie przepływu, jest liczba ekspansji ε, której wartość dla gazów jest mniejsza od 1. Liczbę ekspansji można wyznaczyć na podstawie zależności (3.31).

$$\varepsilon = 1 - \left(0{,}351 + 0{,}256\,\beta^4 + 0{,}93\,\beta^8\right)\left[1 - \left(\frac{p_2}{p_1}\right)^{\frac{1}{\kappa}}\right] \tag{3.31}$$

gdzie p_1 jest ciśnieniem zmierzonym po stronie dopływowej kryzy, p_2 to ciśnienie zmierzone po stronie odpływowej kryzy zaś κ jest wykładnikiem izentropy. Dla powietrza $\kappa = 1{,}4$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

3.10.3. Rozwiązanie

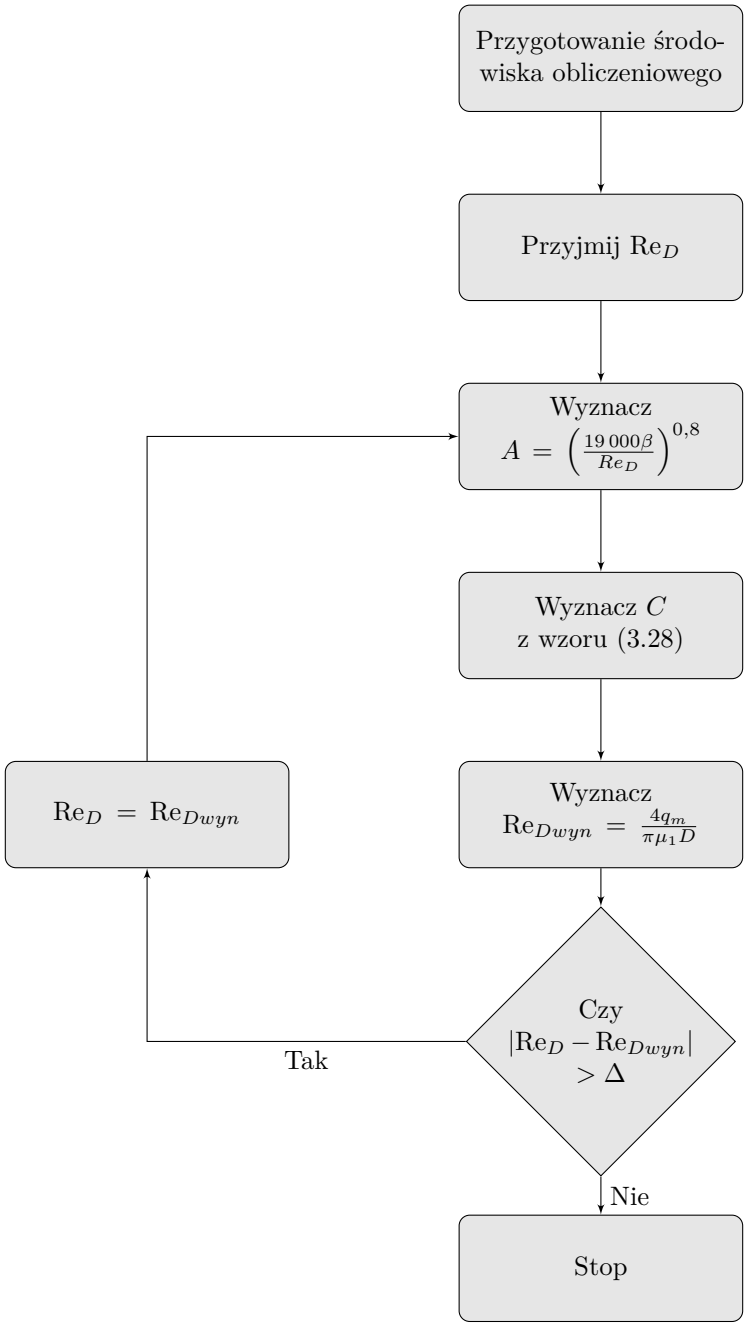
Ponieważ we wzorze (3.25) występuje współczynnik przepływu C, który zależy od liczby Reynoldsa Re, zaś liczbę Reynoldsa wyznacza się na podstawie natężenia przepływu, konieczne jest postępowanie iteracyjne.

Ze względu na niewielką złożoność obliczeniową, rozwiązanie odbywać się będzie za pomocą metody iteracji prostej. Wykorzystanie metody Newtona wiązałoby się z pracochłonnym wyznaczaniem pochodnej $q'_m(\mathrm{Re}_D)$. Zastosowanie metody siecznych okazuje się nieekonomiczne ze względu na większą złożoność obliczeniową i niewielki zysk szybkości rozwiązania. W związku z tym, tok obliczeń przebiega zgodnie z algorytmem przedstawionym na rysunku 3.10. Przyjęta dokładność wyznaczenia liczby Reynoldsa wynosi $\Delta = 1$. Obliczenia zostaną zakończone, gdy różnica między wartościami liczby Reynoldsa, obliczonymi w kolejnych iteracjach, będzie mniejsza niż Δ. Dane wejściowe, wyniki pośrednie oraz wyznaczona wartość liczby Reynoldsa w każdej z przeprowadzonych iteracji zostały zamieszczone w tabeli 3.2

Tabela 3.2. Wyniki kolejnych iteracji

Iteracja	Przyjęte	Obliczenia	Wynik
	$\mathrm{Re}_D = 10\,000$		
1	$\mathrm{Re}_D = 10\,000$	$A = 1{,}06800$	
		$C = 0{,}63084$	
		$q_m = 0{,}02801$	$\mathrm{Re}_{Dwyn} = 27\,728{,}983$
2	$\mathrm{Re}_D = 27\,728{,}983$	$A = 0{,}47231$	
		$C = 0{,}62133$	
		$q_m = 0{,}02759$	$\mathrm{Re}_{Dwyn} = 27\,311{,}032$
3	$\mathrm{Re}_D = 27\,311{,}032$	$A = 0{,}47808$	
		$C = 0{,}62143$	
		$q_m = 0{,}02759$	$\mathrm{Re}_{Dwyn} = 27\,315{,}395$
4	$\mathrm{Re}_D = 27\,315{,}395$	$A = 0{,}47802$	
		$C = 0{,}62143$	
		$q_m = 0{,}02759$	$\mathrm{Re}_{Dwyn} = 27\,315{,}349$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 3.10. Algorytm postępowania przy wyznaczaniu strumienia objętości za pomocą zwężki

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

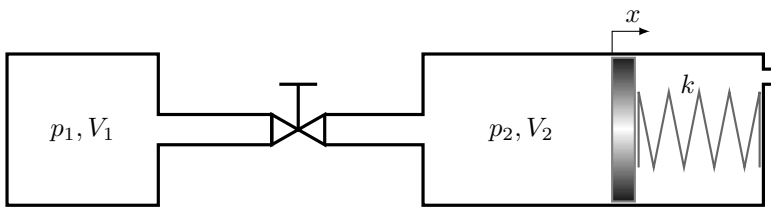
Na podstawie wyznaczonego wydatku masowego q_m można wyliczyć wydatek objętościowy zgodnie ze wzorem (3.26). Obliczony wydatek objętościowy dla aktualnych danych wejściowych wyniesie

$$q_V = 2{,}39322 \cdot 10^{-2}$$

Kolejny przykład dotyczy sytuacji, gdy zbiornik z elastyczną ścianą jest napełniany gazem pod ciśnieniem.

3.10.4. Przykład obliczeniowy

Zbiornik o objętości $V_1 = 0{,}2\ \mathrm{m}^3$ zawierający powietrze pod ciśnieniem $p_1 = 3\ \mathrm{MPa}$ połączono ze zbiornikiem o objętości $V_2 = 1{,}2\ \mathrm{m}^3$ zawierającym powietrze pod ciśnieniem $p_2 = 1$ bar. Temperatura w obu zbiornikach jest równa temperaturze otoczenia i wynosi $T = 300$ K. Wyznaczyć wartość ciśnienia panującego w zbiornikach po ich połączeniu, przyjmując że jedna ściana drugiego zbiornika jest ruchoma, zaś współczynnik sprężystości układu przeciwdziałającego przemieszczeniu ściany wynosi $k = ax^2$. Ciśnienie wyznaczyć po wyrównaniu temperatury z otoczeniem. Wartość współczynnika $a = 10^6$ N/m. Wyznaczyć przemieszczenie ściany zbiornika drugiego, przyjmując że ma postać walca, zaś jego średnica wynosi $d = 1$ m. Powietrze potraktować jako gaz doskonały. Ciśnienie otoczenia wynosi $p_0 = 1$ bar. Schemat układu przedstawiono na rysunku 3.11.



Rysunek 3.11. Schemat analizowanego układu

Rozwiązanie. Należy zapisać równania stanu gazu doskonałego dla obu zbiorników, oraz równanie wyrażające zasadę zachowania masy w analizowanym układzie. Wykorzystując te zależności, można wyznaczyć ciśnienie w zbiornikach, które pozwoli wyznaczyć przemieszczenie ściany.

Masa gazu przed połączeniem wynika z równania stanu gazu doskonałego

$$p_1 V_1 = m_1 RT$$

$$p_2 V_2 = m_2 RT$$

całkowita masa gazu zatem wyniesie

$$m = m_1 + m_2 = \frac{p_1 V_1 + p_2 V_2}{RT}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Powierzchnia przekroju drugiego zbiornika wynosi

$$S = \frac{1}{4}\pi d^2$$

Ponieważ ciśnienia działające na obie strony ruchomej ściany są jednakowe, więc przemieszczenie tej ściany wynosi $x = 0$. Całkowita objętość zbiorników wypełnionych gazem po połączeniu będzie wynosić

$$V = V_1 + V_2 + xS = V_1 + V_2 + x\frac{\pi d^2}{4}$$

Równanie wiążące siłę F i przemieszczenie sprężyny x ma postać $F = kx$. Wobec tego przy panującym w zbiorniku ciśnieniu przemieszczenie x można wyznaczyć z równania

$$x = S\frac{p - p_0}{k} = \frac{\pi d^2}{4}\frac{p - p_0}{ax^2} \quad \Leftrightarrow \quad x^3 = \frac{\pi d^2(p - p_0)}{4a} \tag{3.32}$$

Równanie stanu gazu doskonałego dla układu po połączeniu i po ustaleniu temperatury będzie miało postać następującą

$$pV = mRT \quad \Leftrightarrow \quad p\left(V_1 + V_2 + x\frac{\pi d^2}{4}\right) = \frac{p_1V_1 + p_2V_2}{RT}RT$$

Po wstawieniu zależności (3.32) otrzymuje się następującą postać równania, które należy rozwiązać

$$p\left(V_1 + V_2\right) + p\sqrt[3]{\frac{\pi d^2(p - p_0)}{4a}}\frac{\pi d^2}{4} = p_1V_1 + p_2V_2$$

co po przeniesieniu wyrażenia z prawej strony równania na lewą można zapisać

$$p\left(V_1 + V_2\right) + p\sqrt[3]{\frac{\pi d^2(p - p_0)}{4a}}\frac{\pi d^2}{4} - p_1V_1 - p_2V_2 = 0 \tag{3.33}$$

Po podstawieniu znanych danych równanie przyjmuje postać następującą

$$\frac{866\,208}{119\,536\,999}p\sqrt[3]{p - 100\,000} + \frac{7}{5}p - 720\,000 = 0 \tag{3.34}$$

Rozwiązanie równania. Do rozwiązania równania (3.34) zostanie wykorzystana metoda Newtona. Lewa strona równania wynosi

$$f(p) = \frac{866\,208}{119\,536\,999}p\sqrt[3]{p - 100\,000} + \frac{7}{5}p - 720\,000$$

Pochodna lewej strony równania (3.34) względem ciśnienia p ma następującą postać

$$f'(p) = \frac{288\,736p}{119\,536\,999\left(p - 100\,000\right)^{\frac{2}{3}}} + \frac{866\,208\left(p - 100\,000\right)^{\frac{1}{3}}}{119\,536\,999} + \frac{7}{5}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zastosowanie metody Newtona do rozwiązania tego równania prowadzi do następującej zależności iteracyjnej

$$p_{i+1} = p_i - \frac{f(p_i)}{f'(p_i)}$$

Niech punktem startowym będzie $p_0 = 200\,000$. Poniżej zostaną zestawione kolejne wyniki.

i	x_i	$\lvert p_i - p_{i-1}\rvert$
0	$p_0 = 200\,000$	
1	$p_1 = 390\,112{,}792$	$190\,112{,}792$
2	$p_2 = 383\,764{,}709$	$6\,348{,}083$
3	$p_3 = 383\,758{,}828$	$5{,}8812$
4	$p_4 = 383\,758{,}828$	$5{,}0819 \cdot 10^{-6}$

Zatem ciśnienie wewnątrz zbiorników po ich połączeniu i wyrównaniu się temperatury z temperaturą otoczenia wyniesie

$$p = p_4 = 383\,758{,}828 \text{ Pa}$$

Przemieszczenie ścianki zbiornika x wyniesie

$$x = \sqrt[3]{\frac{\pi d^2 (p - p_0)}{4a}} = \sqrt[3]{\frac{\pi \cdot 1^2 \cdot (383\,758{,}828 - 100\,000)}{4 \cdot 10^6}} = 0{,}606 \text{ m}$$

Do rozwiązania wyprowadzonego równania (3.34) wykorzystano metodę Newtona, ponieważ stosunkowo łatwo można było wyznaczyć pochodną. Jeżeli wyznaczenie pochodnej lewej strony równania jest zbyt złożone, można użyć metody siecznych.

3.11. Miejsca zerowe wielomianów

Aby znaleźć miejsca zerowe wielomianów, można wykorzystać w zasadzie dowolne metody, chociażby metodę siecznych czy metodę Newtona, jeśli oczywiście rozwiązywany problem spełnia kryteria zbieżności. W niniejszym podrozdziale omówiona zostanie idea znajdowania rzeczywistych pojedynczych pierwiastków wielomianu $W(x)$.

Używając, na przykład, metody Newtona, należy obliczyć jeden z pierwiastków x_0^*. Następnie dzieli się wielomian $W(x)$ przez dwumian $(x - x_0^*)$. Kolejnym krokiem jest wyznaczenie pierwiastka otrzymanego wielomianu. Wykorzystując otrzymany pierwiastek jako punkt startowy, należy powtórzyć obliczenia dla oryginalnego wielomianu i dopiero tak uzyskany wynik przyjąć jako wartość pierwiastka. Taką procedurę powtarza się aż do znalezienia wszystkich pierwiastków wielomianu. Ponieważ

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wartości pierwiastków wielomianu są obarczone błędem, zatem pomimo starań może dojść do tego, że w kolejnym kroku wyznacza się pierwiastek innego wielomianu niż w obliczeniach z dokładną wartością pierwiastka. Opisany sposób jest na tyle uniwersalny, że można zastosować dowolną metodę rozwiązywania równań nieliniowych. W niektórych przypadkach mogą wystąpić problemy ze zbieżnością, ale jest to cena prostoty algorytmu.

Inną, dosyć ciekawą propozycję stanowi metoda mająca własność dobrej zbieżności globalnej, zwana metodą Laguerre'a, która zostanie teraz skrótowo przedstawiona.

3.11.1. Metoda iteracyjna Laguerre'a

Metoda iteracyjna Laguerre'a [5] jest wykorzystywana do wyznaczania wszystkich pierwiastków równania w postaci

$$p(x) = 0$$

co można zapisać po rozwinięciu jako

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n = 0, \qquad a_0 \neq 0$$

Można ją zapisać w postaci następujących zależności

$$x_{i+1} = x_i - \frac{np(x_i)}{p'(x_i) \pm \sqrt{H(x_i)}} \tag{3.35}$$

gdzie

$$H(x) = (n-1)[(n-1)(p'(x))^2 - np(x)p''(x)]$$

natomiast n jest stopniem wielomianu $p(x)$. Znak w mianowniku zależności (3.35) wybiera się tak, by różnica $|x_{i+1} - x_i|$ była najmniejsza. Niestety, w każdym kroku konieczne jest wyznaczenie wartości $p'(z_i)$, $p''(z_i)$ oraz $p(z_i)$, co powoduje, że metoda ta wymaga znacznego nakładu obliczeniowego.

Warto pamiętać, że w przypadku wielokrotnych pierwiastków wielomianu zeruje się nie tylko sam wielomian, ale i jego pochodna. Wobec tego pojawiają się trudności w stosowaniu niektórych metod.

3.12. Rozwiązywanie równań nieliniowych z wykorzystaniem MATLAB-a

Przedstawiony materiał pozwala zrozumieć podstawowe metody służące do rozwiązywania algebraicznych równań nieliniowych. Od inżyniera mechanika zazwyczaj nie wymaga się pisania programów autorskich rozwiązujących takie równania, tym bardziej że napisanie dobrego programu wymaga długiego testowania i skrupulatnego sprawdzania jego jakości. Z tego powodu poniżej zostaną skrótowo przedstawione sposoby rozwiązywania takich problemów za pomocą MATLAB-a lub też wolnodostępnego oprogramowania Octave.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Do rozwiązywania takich równań może służyć funkcja `fzero`, do znajdowania zer wielomianów funkcja `roots`. Za pomocą MATLAB-a czy Octave można też rozwiązywać równania symbolicznie.

Aby wykorzystać funkcję `fzero` należy:

— zdefiniować rozwiązywane równanie np. w postaci funkcji `inline`,
— użyć funkcji `fzero`,

czyli

```
1 f='x-cos(x)'
  rown= inline (f,'x')
3 fzero(rown,1)
```

Aby zrozumieć użycie funkcji `roots`, niezbędna jest wiedza na temat sposobu zapisu wielomianów w MATLAB-ie czy Octave. Są one zapisywane w postaci wektora współczynników, czyli wielomian

$$W = a_0 + a_1 x + a_2 x^2 \ldots + a_n x^n$$

zapisuje się w postaci `W=[`a_n, ... a_2, a_1, a_0`]`. Wykorzystanie funkcji `roots` może wyglądać następująco

```
1 % zera wielomianu W=x^3-3
  a=[1 0 0 -3]
3 roots(a)
```

Funkcja `roots` zwraca wartości wszystkich pierwiastków wielomianu.

Symboliczne rozwiązywanie równań w MATLAB-ie realizuje się dosyć prosto, niemniej tematyka obliczeń symbolicznych została zgrupowana w oddzielny rozdział. Obliczenia symboliczne za pomocą pakietów algebry komputerowej (CAS) zostały skrótowo opisane w rozdziale 12.

3.13. Podsumowanie

Trudno jest jednoznacznie wskazać najlepszą metodę rozwiązywania algebraicznych równań nieliniowych. Wybór zależy od posiadanej a priori wiedzy na temat problemu. Liczba pierwiastków, ich krotność oraz wstępne oszacowanie ich położenia na osi, pozwalają na wybór właściwej metody obliczeniowej.

W praktyce codziennej inżynier, a szczególnie mechanik, często nie przywiązuje wagi do wyboru metody, rzadko pisze samodzielnie programy do rozwiązywania problemów. Najczęściej wykorzystuje gotowe pakiety numeryczne – komercyjne i zamknięte, jak MATLAB czy Mathematica, albo otwarte jak GNU Octave czy Maxima. Takie praktyczne podejście nie zwalnia inżyniera z odpowiedzialności za wyniki obliczeń. Z tego też powodu należy zawsze sprawdzać wyniki obliczeń chociażby w najprostszy sposób – podstawiając wartości do rozwiązywanych równań.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 4

Całkowanie numeryczne

4.1. Wstęp

Jednym z częściej występujących działań przy rozwiązywaniu problemów inżynierskich jest całkowanie. Pod pojęciem całkowania numerycznego rozumie się wyznaczanie wartości całki oznaczonej

$$\int_a^b f(x)\mathrm{d}x \approx \widehat{R}(h) \tag{4.1}$$

co zazwyczaj wiąże się z określeniem kroku całkowania definiującego węzły, w których wyznacza się wartości funkcji podcałkowej. $\widehat{R}(h)$ oznacza przybliżoną wartość całki, obliczoną przy kroku całkowania h. Zazwyczaj problem całkowania rozwiązywany jest za pomocą interpolacji, co oznacza, że interpoluje się funkcję podcałkową np. wielomianem, a następnie dokonuje się całkowania uzyskanej funkcji interpolującej, np. wielomianu.

Wydaje się, iż najłatwiejsze pojęciowo jest geometryczne podejście do problemu całkowania numerycznego. Zgodnie z interpretacją geometryczną przedstawioną na rysunku 4.1, wartość całki oznaczonej jest równa polu ograniczonemu krzywą $y = f(x)$, osią OX oraz prostymi $x = a$ oraz $x = b$.

Tak sformułowany problem można rozwiązać na wiele sposobów, spośród których autorzy wybrali kilka poniżej wymienionych. Przedstawione metody są stosunkowo proste i stanowią demonstrację sposobu postępowania przy całkowaniu:

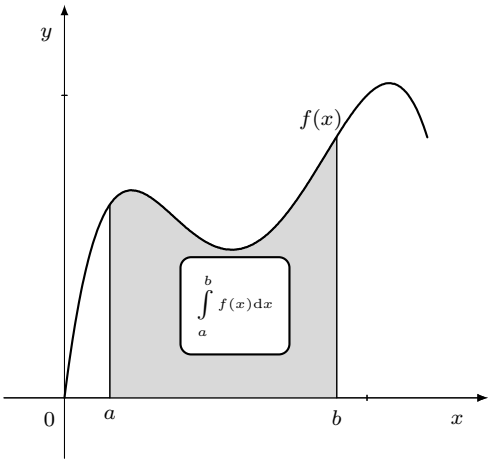
— kwadratury Newtona–Cotesa, a w szczególności:
  - metoda trapezów,
  - metoda Simpsona,

— kwadratury Gaussa.

Konstrukcja tych metod sprowadza się do podziału pola pod krzywą $y = f(x)$ na pola o kształcie prostokąta, trapezu czy trapezu krzywoliniowego, co jest równoważne interpolacji funkcji podcałkowej wielomianami zerowego, pierwszego, drugiego i wyższych stopni.

Niniejsza publikacja aplikuje opisane metody całkowania numerycznego do problemów jednowymiarowych, czyli takich w których pod całką występuje funkcja jednej zmiennej. Nic nie stoi na przeszkodzie by opisywane postępowanie uogólnić i zastosować do zagadnień dwu- i więcejwymiarowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.1. Interpretacja geometryczna całki oznaczonej

4.2. Metoda prostokątów

Jeżeli w problemie (4.1) dokonana zostanie aproksymacja funkcji $f(x)$ za pomocą prostokątów, to otrzyma się najbardziej oczywiste rozwiązanie problemu, czyli aproksymuje się funkcję podcałkową funkcją stałą. Wysokość prostokąta można wyznaczać z lewej strony przedziału, z prawej strony przedziału czy też w środku (rys. 4.2). Najdokładniejszy wynik otrzymuje się przy wyborze wysokości prostokąta na podstawie wartości funkcji $f(x)$ w środku przedziału.

W wyniku podziału przedziału $\langle a, b\rangle$ na n podprzedziałów o długości h wyznaczonych przez punkty

$$\begin{aligned} x_0 &= a \\ x_i &= x_0 + ih \\ x_n &= b \end{aligned} \tag{4.2}$$

gdzie $h = \frac{b-a}{n}$, otrzymuje się wysokości prostokątów w punktach leżących w środku podprzedziałów, jako odpowiednie wartości funkcji

$$f\left(x_i + \frac{h}{2}\right)$$

Przybliżoną wartość całki oznaczonej można wyznaczyć jako wartość następującego wyrażenia

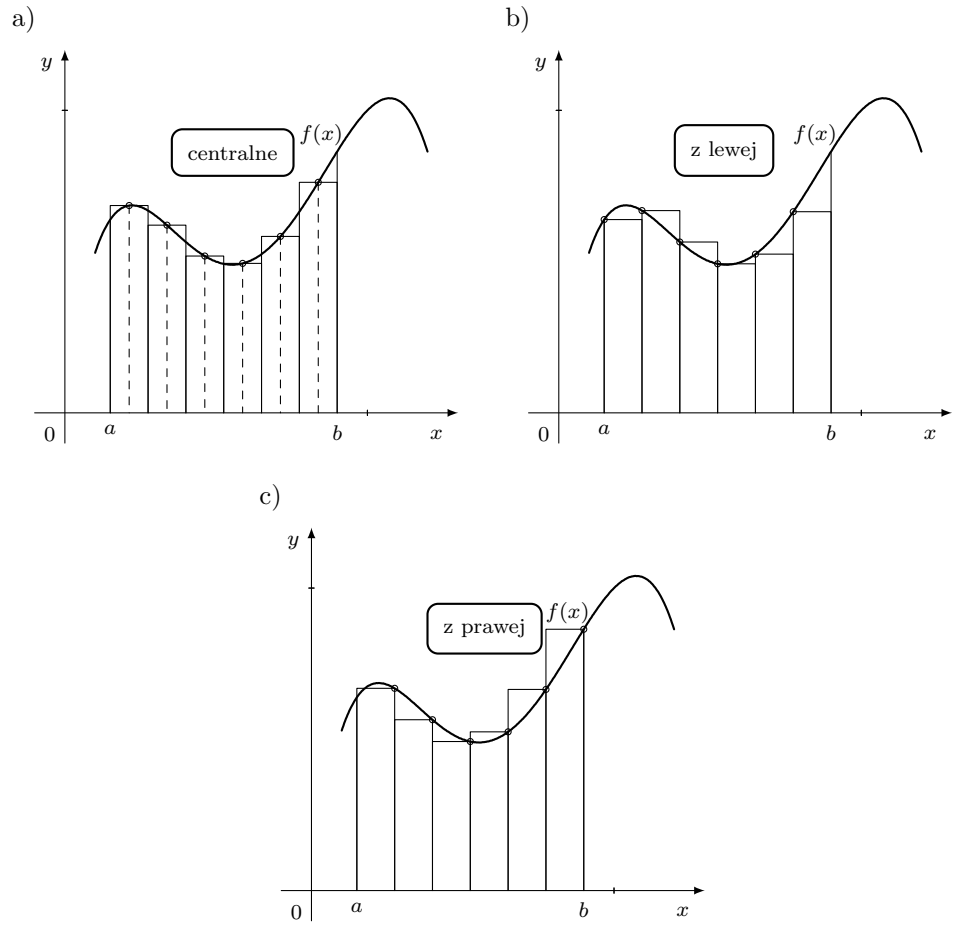
$$\int_a^b f(x)\mathrm{d}x \approx \widehat{R}(h) = h\sum_{i=0}^{n-1} f\left(x_i + \frac{h}{2}\right) = h\sum_{i=1}^{n} f\left(x_i - \frac{h}{2}\right) \tag{4.3}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Dla prostokątów, których wysokości wyznaczają wartości funkcji na lewej lub prawej granicy przedziałów, odpowiednie wzory wyglądają następująco

$$\int_a^b f(x)\mathrm{d}x \approx \widehat{R}(h) = h\sum_{i=1}^{n} f(x_{i-1}) = h\sum_{i=0}^{n-1} f(x_i) \tag{4.4}$$

$$\int_a^b f(x)\mathrm{d}x \approx \widehat{R}(h) = h\sum_{i=1}^{n} f(x_i) \tag{4.5}$$



Rysunek 4.2. Różne sposoby konstruowania prostokątów przy całkowaniu metodą prostokątów – centralne a), z lewej b) i z prawej c)

Wyniki całkowania metodą prostokątów są obarczone błędem obcięcia (związanym z ograniczeniem rozwinięcia funkcji w szereg do pierwszych wyrazów) proporcjo-

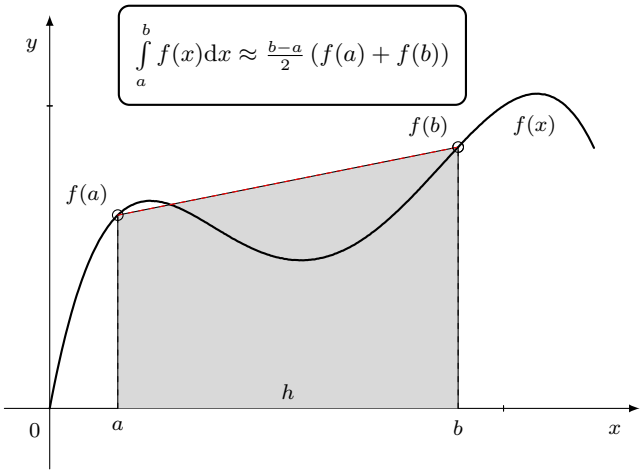
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

nalnym do kwadratu kroku. Przy automatycznym doborze kroku całkowania można wykorzystywać poprzednie wyniki, by zmniejszyć liczbę koniecznych obliczeń. Wyniki otrzymane metodą prostokątów z wartościami w środku przedziałów są dokładniejsze niż z metody trapezów.

4.3. Metoda trapezów

Jeżeli zamiast aproksymacji przebiegu funkcji $f(x)$ funkcjami stałymi zostanie zastosowana aproksymacja funkcjami liniowymi, to w konsekwencji otrzyma się metodę trapezów. Wykorzystując interpolację Lagrange'a (patrz podrozdział 2.1.6), można napisać, że funkcja podcałkowa jest równa $f(a)l_a + f(b)l_b$, gdzie l_a oznacza wielomian Lagrange'a dla węzła a i podobnie l_b dla węzła b. Czyli tok postępowania jest zgodny z poniższymi przekształceniami

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b f(a)l_a + f(b)l_b\,\mathrm{d}x = \int_a^b f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}\mathrm{d}x = \\ = \frac{b-a}{2}\left(f(a)+f(b)\right) = \frac{1}{2}\left(f(a)+f(b)\right)h \quad (4.6)$$



Rysunek 4.3. Ilustracja całkowania metodą trapezów

Rysunek 4.3 jest ilustracją przekształceń (4.6). Aby zwiększyć dokładność, podobnie jak w metodzie prostokątów, przyjmuje się dyskretyzację przedziału $\langle a, b\rangle$, zgodnie z równaniami (4.2), na n podprzedziałów o długości h każdy. Następnie buduje się trapezy o wysokości h (wzdłuż osi OX), długości podstaw stanowią wartości funkcji w węzłach ograniczających podprzedział. Obliczając pola kolejnych trapezów oraz

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

sumując je, wyznacza się przybliżoną wartość całki. Zgodnie z rysunkiem 4.4 pole powierzchni P_i trapezu o indeksie i wyraża się wzorem

$$P_i = \frac{1}{2}\left(f(x_{i-1}) + f(x_i)\right)h \tag{4.7}$$

Wartość całki będzie sumą wszystkich trapezów, co prowadzi do zależności

$$\int_a^b f(x)\mathrm{d}x \approx \widehat{T}(h) = \sum_{i=1}^{n} P_i = \frac{h}{2}\sum_{i=1}^{n}(f(x_{i-1}) + f(x_i)) \tag{4.8}$$

Sumę występującą w tej zależności można rozbić i napisać

$$\sum_{i=1}^{n}(f(x_{i-1}) + f(x_i)) = \sum_{i=1}^{n} f(x_{i-1}) + \sum_{i=1}^{n} f(x_i) =$$
$$f(x_0) + \sum_{i=1}^{n-1} f(x_i) + f(x_n) + \sum_{i=1}^{n-1} f(x_i) = f(x_0) + f(x_n) + 2\sum_{i=1}^{n-1} f(x_i) \tag{4.9}$$

Zatem całkowanie metodą trapezów można zapisać za pomocą następującej formuły (4.10)

$$\widehat{T}(h) = \frac{h}{2}\left[f(x_0) + f(x_n) + 2\sum_{i=1}^{n-1} f(x_i)\right] \tag{4.10}$$

lub

$$\widehat{T}(h) = h\left[\frac{f(x_0) + f(x_n)}{2} + \sum_{i=1}^{n-1} f(x_i)\right] \tag{4.11}$$

przy czym ta druga (4.11) wymaga wykonania jednego mnożenia mniej.

Warto zdawać sobie sprawę, iż jeżeli precyzja obliczeń wynosi ε, to błędy popełniane podczas całkowania będą spełniać warunek

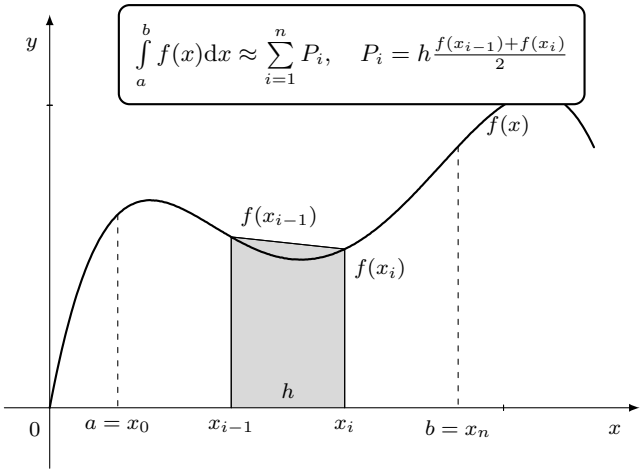
$$\delta \leqslant hn\varepsilon = (b-a)\varepsilon$$

Chociaż metoda trapezów przy niewielkim kroku daje dobre wyniki, to niewielkim nakładem obliczeniowym można ją zdecydowanie poprawić. Poniżej przedstawiona zostanie ekstrapolacja Richardsona w metodzie trapezów, pozwalająca zmniejszać błąd metody – teoretycznie do dowolnie małej wartości. Bardziej szczegółowo zostanie omówiony ten temat w podrozdziale 4.4 dotyczącym metody Romberga.

4.3.1. Ekstrapolacja Richardsona w metodzie trapezów

Wykorzystując fakt, iż w metodzie trapezów błąd jest w przybliżeniu proporcjonalny do kwadratu kroku, można podjąć próbę wyznaczenia dokładniejszej wartości całki oznaczonej przez eliminację przynajmniej części tego błędu. W tym celu wyznacza się dwukrotnie wartość całki.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.4. Konstruowanie trapezu w metodzie trapezów

Za pierwszym razem wyznacza się całkę z krokiem o długości h, za drugim – z krokiem o długości będącej wielokrotnością h, na przykład równym $2h$. Można zatem napisać, że wartość dokładna całki oznaczonej jest sumą wartości wyznaczonej metodą trapezów oraz błędu proporcjonalnego do kwadratu kroku całkowania h^2

$$\begin{aligned} I &\approx T(h) + kh^2 \\ I &\approx T(2h) + k(2h)^2 \end{aligned}$$

Zakłada się przy tym, że przy całkowaniu z krokiem dwa razy większym stała błędu całkowania pozostaje taka sama. Ponieważ na wartość stałej błędu k wpływa wartość pierwszej pochodnej funkcji podcałkowej, można przyjąć, że k jest stałe.

Następnie wyznaczając składnik oznaczający błąd z pierwszego równania i wstawiając do drugiego, można uzyskać równanie w postaci

$$I \approx T(2h) + 4\,(I - T(h)),$$

co prowadzi do zależności

$$I \approx T(h) + \frac{1}{3}(T(h) - T(2h)) \tag{4.12}$$

Warto zwrócić uwagę, że w ogólnym przypadku błąd w metodzie może być proporcjonalny do h^n. Krok w kolejnych przybliżeniach przyjmuje się najczęściej h oraz $2h$, a wyniki obliczeń z krokiem h oraz $2h$ oznaczy się odpowiednio jako $M(h)$ i $M(2h)$. Przeprowadzając rozumowanie podobne do powyżej przedstawionego, można otrzymać ogólną postać wzoru ekstrapolacyjnego Richardsona taką jak w równaniu (4.13).

$$I \approx M(h) + \frac{1}{2^n - 1}\,(M(h) - M(2h)) \tag{4.13}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wartość całki oznaczonej obliczona za pomocą ekstrapolacji Richardsona jest znacznie dokładniejsza niż cząstkowe przybliżenia metody trapezów, ponieważ zostaje usunięta główna składowa błędu całkowania, która jest proporcjonalna do h^2. Warto zwrócić uwagę na fakt, że metoda trapezów z ekstrapolacją Richardsona daje wyniki identyczne z opisaną dalej metodą Simpsona. Jak łatwo się przekonać, można otrzymać metodę Simpsona z metody trapezów z ekstrapolacją Richardsona. Dla trzech równoodległych węzłów x_0, x_1 oraz x_2 znane są trzy wartości funkcji $f(x_0)$, $f(x_1)$ i $f(x_2)$. Metoda trapezów dla tego przypadku będzie miała postać

$$T(h) = h\left(\frac{f(x_0)+f(x_1)}{2} + \frac{f(x_1)+f(x_2)}{2}\right) = \frac{h}{2}\left(f(x_0) + 2\,f(x_1) + f(x_2)\right)$$

$$T(2h) = \frac{2\,h}{2}\left(f(x_0) + f(x_2)\right) = h\left(f(x_0) + f(x_2)\right)$$

Ekstrapolacja Richardsona w metodzie trapezów w tym przypadku przyjmie następującą postać

$$\begin{aligned} I \approx T(h) + \frac{1}{3}\left(T(h) - T(2\,h)\right) &= \frac{h}{2}\left(f(x_0) + 2\,f(x_1) + f(x_2)\right) + \\ &+ \frac{1}{3}\left(\frac{h}{2}\left(f(x_0) + 2\,f(x_1) + f(x_2)\right) - h\left(f(x_0) + f(x_2)\right)\right) = \\ = \frac{h}{6}\left(3\,f(x_0) + 6\,f(x_1) + 3\,f(x_2) + f(x_0) + 2\,f(x_1) + f(x_2) - 2\,f(x_0) - 2\,f(x_2)\right) &= \\ = \frac{h}{6}\left(2\,f(x_0) + 8\,f(x_1) + 2\,f(x_2)\right) = \frac{h}{3}\left(f(x_0) + 4\,f(x_1) + f(x_2)\right) & \quad (4.14) \end{aligned}$$

Zatem metoda trapezów z ekstrapolacją Richardsona ma bardzo korzystne właściwości. Metoda Simpsona wymaga jednak mniejszej liczby działań, jest zatem bardziej ekonomiczna obliczeniowo.

4.4. Metoda Romberga

Metoda Romberga stanowi przykład ciekawego rozwinięcia zastosowania metody ekstrapolacji Richardsona do metody trapezów. Sama metoda wynika wprost ze wzoru sumacyjnego Eulera–Maclaurina [5], na podstawie którego stwierdza się, że błąd wyznaczenia wartości całki oznaczonej za pomocą metody trapezów wyraża się następującą zależnością

$$\Delta = T(h) - \int_a^b f(x)\mathrm{d}x = c_1 h^2 + c_2 h^4 + c_3 h^6 + \cdots \qquad (4.15)$$

co pozwala zastosować wielokrotnie ekstrapolację Richardsona w postaci (4.13) do wyników, które uprzednio pozbawiono składników błędu proporcjonalnego do niższych potęg h. Wygodnie jest to zrobić, posługując się tablicą zawierającą w wierszach

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wyniki obliczeń wartości całki metodą trapezów z coraz mniejszym krokiem, zaś w kolumnach kolejne rzędy ekstrapolacji Richardsona. Krok za każdym razem zmniejsza się dwukrotnie. Ponieważ w kolejnych wierszach można wykorzystać do ekstrapolacji wyniki z wcześniejszych wierszy, tablica ma kształt trójkątny. Poniżej zilustrowano opisaną ideę.

$T(\frac{h}{n})$	$\frac{1}{3}\Delta$	T_{ij}	$\frac{1}{15}\Delta_2$	T_{ijk}
$T(h)$				
$T(\frac{h}{2})$	$\frac{1}{3}(T(\frac{h}{2}) - T(h))$	$T_{12} = T(\frac{h}{2}) + \frac{1}{3}\Delta$		
$T(\frac{h}{4})$	$\frac{1}{3}(T(\frac{h}{4}) - T(\frac{h}{2}))$	$T_{23} = T(\frac{h}{4}) + \frac{1}{3}\Delta$	$\frac{1}{15}(T_{23} - T_{12})$	$T_{123} = T_{23} + \frac{1}{15}\Delta$

Aby uprościć programowanie tej metody, powyższą tablicę można zapisać nieco inaczej za pomocą następujących oznaczeń

$$\begin{array}{llllll} R(0,0) & & & & & \\ R(1,0) & R(1,1) & & & & \\ R(2,0) & R(2,1) & R(2,2) & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ R(M,0) & R(M,1) & R(M,2) & \ldots & R(M,M) \end{array}$$

Gdzie pierwsza kolumna zawiera wartości całki wyznaczone metodą trapezów. Pierwszy wiersz to wartość całki obliczona z najdłuższym krokiem, każdy następny wyznaczony przy kroku zmniejszonym o połowę. Następnie poszczególne składniki $R(i,j)$ można wyliczyć zgodnie z wzorem

$$R(i,j) = R(i,j-1) + \frac{1}{4^j - 1}\left[R(i,j-1) - R(i-1,j-1)\right] \tag{4.16}$$

Oczywiście obliczenia można kontynuować dla kolejnych długości kroku, co pozwoli uzyskać wynik całkowania w zasadzie z dowolną dokładnością, oczywiście ograniczoną przez dokładność maszynową. Warto zauważyć, iż odpowiednia wartość poprawki $\frac{1}{2^n - 1}\Delta$ pozwala oszacować dokładność wyznaczenia wartości całki.

4.4.1. Przykład obliczeniowy

Sformułowanie problemu. Metodą Romberga wyznaczyć wartość całki z funkcji $\sin x$ z dokładnością $\Delta < 10^{-6}$

$$\int_0^{\frac{\pi}{2}} \sin x \, \mathrm{d}x$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Należy wyznaczyć wartość całki metodą trapezów dla kroków o długości h, $\frac{1}{2}h$, $\frac{1}{4}h$ oraz $\frac{1}{8}h$. Poniższa tablica ilustruje przeprowadzone obliczenia. Pierwsza kolumna n zawiera liczbę podprzedziałów, na którą jest dzielony cały przedział całkowania.

n	$T(\frac{\pi}{2n})$	$T+1/3\Delta$	$T+1/15\Delta_2$	$T+1/63\Delta_3$
1	0,785398163			
2	0,948059449	1,002279877		
4	0,987115801	1,000134585	0,999991565	
8	0,996785172	1,000008296	0,999999876	1,000000008

Wartość całki wyznaczona analitycznie wynosi 1. Oszacowaniem błędu jest wartość poprawki wynikająca z wartości w czwartej kolumnie, która wynosi $\frac{1}{63}\Delta_3$, czyli

$$\frac{1}{63} \cdot (0{,}999999876 - 0{,}999991565) = 1{,}32 \cdot 10^{-7}$$

Zatem wymagana dokładność została osiągnięta. Błąd wyznaczenia wartości całki wynosi $8 \cdot 10^{-9}$, zaś oszacowany błąd jest dwa rzędy większy. Jest to zatem bardzo bezpieczne oszacowanie wielkości błędu popełnianego przy całkowaniu.

Warto zwrócić uwagę, iż do obliczeń należało wyznaczyć 9 wartości funkcji podcałkowej, a następnie przeprowadzać obliczenia na tych wartościach. Co więcej, wzór (4.11) opisujący metodę trapezów ma postać pozwalającą na pewną optymalizację procedur numerycznych, przez co metoda Romberga staje się bardzo wydajną metodą całkowania numerycznego przy stałej długości kroku.

4.5. Wzór Simpsona

Wzór Simpsona powstaje, gdy funkcja podcałkowa (4.1) zostanie aproksymowana wielomianem stopnia drugiego, którego wykresem jest parabola. Aproksymację stosuje się w przedziałach całkowania. W literaturze można spotkać różne podejścia do podziału – wymaga się, by liczba przedziałów była parzysta lub dokonuje się podziałów dodatkowych, wyznaczając wartości w środku przedziału. W związku z tym można spotkać różne postacie wzorów.

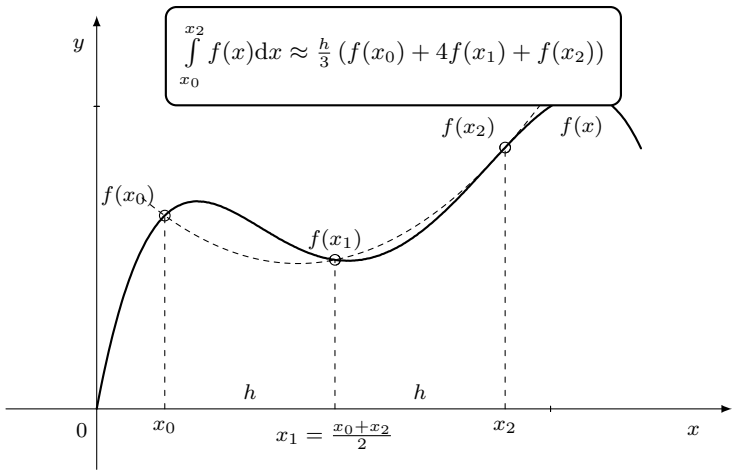
Poniżej zostanie przedstawione elementarne wyprowadzenie metody Simpsona dla jednego przedziału całkowania, co nie umniejsza ogólności wyprowadzenia. Niewielkim nakładem sił można wyprowadzić złożony wzór na całkowanie przedziałowe metodą Simpsona. Ilustrację wyprowadzenia stanowi rysunek 4.5.

Niech będzie dany przedział $\langle x_0, x_2\rangle$ o długości $2\,h$, w środku przedziału leży wartość $x_1 = x_0 + h = x_2 - h$. Wyznacza się pole ograniczone parabolą przechodzącą przez punkty: $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$ i osią OX. Parabola interpolująca

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

funkcję podcałkową ma wzór $W(x) = ax^2 + bx + c$, w związku z tym wartości funkcji dla węzłów x_0, x_1, x_2 są następujące

$$\begin{cases} f(x_0) &= ax_0^2 + bx_0 + c \\ f(x_1) &= ax_1^2 + bx_1 + c \\ f(x_2) &= ax_2^2 + bx_2 + c \end{cases} \tag{4.17}$$



Rysunek 4.5. Ilustracja całkowania metodą Simpsona

Wartość całki oznaczonej wyniesie

$$\int_{x_0}^{x_2} f(x)\,\mathrm{d}x \approx \int_{x_0}^{x_2} W(x)\,\mathrm{d}x = \int_{x_0}^{x_2} \left(ax^2 + bx + c\right)\,\mathrm{d}x = \frac{a}{3}x^3 + \frac{b}{2}x^2 + cx\Big|_{x_0}^{x_2} =$$

$$= \frac{a}{3}(x_2^3 - x_0^3) + \frac{b}{2}(x_2^2 - x_0^2) + c(x_2 - x_0) \tag{4.18}$$

Po zastosowaniu wzorów skróconego mnożenia

$$x_2^3 - x_0^3 = (x_2 - x_0)\left(x_2^2 + x_2x_0 + x_0^2\right)$$

oraz

$$x_2^2 - x_0^2 = (x_2 - x_0)(x_2 + x_0)$$

i wyłączeniu przed nawias istotnych czynników można uzyskać następującą zależność

$$\int_{x_0}^{x_2} W(x)\mathrm{d}x = \frac{x_2 - x_0}{6}\left[2a(x_2^2 + x_2x_0 + x_0^2) + 3b(x_2 + x_0) + 6c\right] =$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$= \frac{x_2 - x_0}{6}\left[2ax_2^2 + 2ax_2x_0 + 2ax_0^2 + 3bx_2 + 3bx_0 + 6c\right] =$$

$$= \frac{h}{3}\left[\underbrace{ax_2^2 + bx_2 + c}_{f(x_2)} + \underbrace{ax_0^2 + bx_0 + c}_{f(x_0)} + 4\underbrace{\left(a\left(\frac{x_2 + x_0}{2}\right)^2 + b\frac{x_2 + x_0}{2} + c\right)}_{f(\frac{x_2+x_0}{2}) = f(x_1)}\right] =$$

$$= \frac{h}{3}\left[f(x_2) + f(x_0) + 4f(x_1)\right] \quad (4.19)$$

Zatem końcowa zależność dla metody Simpsona wygląda tak

$$\int_{x_0}^{x_2} f(x)\mathrm{d}x \approx \widehat{S}(h) = \frac{1}{3}h(f(x_0) + 4f(x_1) + f(x_2)) \quad (4.20)$$

Nie jest to jedyny sposób wyprowadzenia tego wzoru. Można też wykorzystać na przykład metodę współczynników nieoznaczonych czy, jak to zostało pokazane w zależności (4.14), z metody trapezów z ekstrapolacją Richardsona.

W praktycznych zagadnieniach przedział całkowania $\langle a, b\rangle$ dzieli się na m podprzedziałów z krokiem $2h$, a następnie realizuje się całkowanie w każdym z podprzedziałów. Takie postępowanie powoduje, że złożony wzór Simpsona przyjmuje postać (4.21).

$$\int_{x_0}^{x_n} f(x)\mathrm{d}x \approx \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{\frac{n}{2}} f(x_{2i-1}) + 2\sum_{i=1}^{\frac{n}{2}-1} f(x_{2i}) + f(x_n)\right) \quad (4.21)$$

Warto odpowiedzieć na pytanie, jak wraz z krokiem zmienia się błąd w metodzie Simpsona. Aby ocenić wielkość błędu całkowania, należy rozwinąć funkcję podcałkową w szereg Taylora. Gdy przyjmie się $x = 0$ na środku przedziału oraz $x = h$, można zapisać

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \mathcal{O}(h^4)$$

Oczywiste jest, że dla wielomianów do trzeciego stopnia włącznie, reszta powyższego wzoru jest zerem. Zatem przyjmując $a = -h$ zaś $b = h$, można zapisać całkę

$$\int_a^b f(x)\mathrm{d}x = \int_{-h}^{h} f(x)\mathrm{d}x = f(0)\int_{-h}^{h}\mathrm{d}x + f'(0)\int_{-h}^{h} x\mathrm{d}x + \frac{f''(0)}{2}\int_{-h}^{h} x^2\mathrm{d}x +$$

$$+ \frac{f''(0)}{6}\int_{-h}^{h} x^3\mathrm{d}x + \mathcal{O}(h^5) = 2hf(0) + 0 + \frac{h^3}{3}f''(0) + 0 + \mathcal{O}(h^5) \quad (4.22)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwijając odpowiednie wyrażenia w szereg Taylora i odejmując stronami, można wykazać, iż w wyrażeniu

$$f''(0) = \frac{1}{h^2}\left(f(-h) - 2f(0) + f(h)\right) + \mathcal{O}(h^2)$$

reszta jest zerem dla wszystkich wielomianów rzędu trzeciego [5]. Ponieważ można udowodnić, iż dla wielomianów stopnia drugiego zachodzi równość

$$f''(a) = \frac{f(a+h) - 2f(a) + f(a-h)}{h^2}$$

otrzymuje się następującą postać wzoru Simpsona

$$\int_{-h}^{h} f(x)\mathrm{d}x = 2hf(0) + \frac{1}{3}h(f(-h) - 2f(0) + f(h)) + \mathcal{O}(h^5) =$$

$$= \frac{h}{3}(f(-h) + 4f(0) + f(h)) + \mathcal{O}(h^5) \quad (4.23)$$

gdzie reszta jest zerem dla wszystkich wielomianów stopnia trzeciego.

Podobnie jak metoda trapezów, metoda ta jest szczególnym przypadkiem kwadratury Newtona–Cotesa. Uzyskiwane wyniki całkowania dla funkcji podcałkowych w postaci wielomianów do trzeciego stopnia włącznie są dokładne. Jak pokazano, daje wyniki identyczne z wynikami metody trapezów z ekstrapolacją Richardsona, ale jest ekonomiczniejsza obliczeniowo.

4.5.1. Przykład obliczeniowy

Sformułowanie problemu. Wyznaczyć wartość całki oznaczonej

$$I = \int_{0}^{\frac{\pi}{2}} \sin x\mathrm{d}x$$

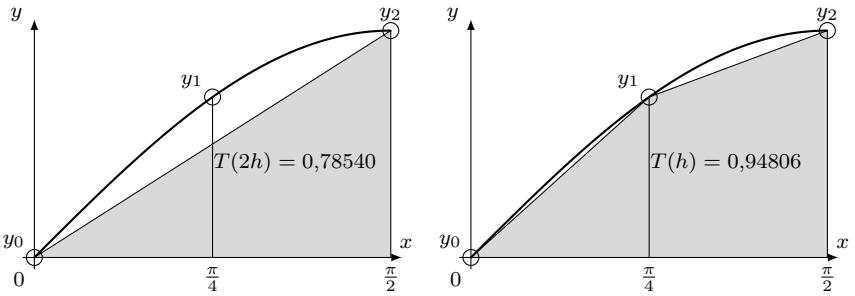
posługując się metodą trapezów z ekstrapolacją Richardsona oraz metodą Simpsona. Do obliczeń użyć możliwie największego kroku całkowania. Wyznaczyć błąd obliczenia wartości całki dla każdej z metod.

Rozwiązanie Krok całkowania w metodzie trapezów z ekstrapolacją Richardsona będzie wynosił: $h = \pi/4$ oraz $2h = \pi/2$. Wartości funkcji podcałkowej dla wybranych kroków znajdują się w tablicy na rysunku 4.6. Wspomniany rysunek przedstawia również sposób konstruowania trapezów do metody trapezów, do całkowania z krokiem $2h$ oraz h. Ostatni z rysunków przedstawia wielomian drugiego stopnia interpolujący funkcję sinus do metody Simpsona.

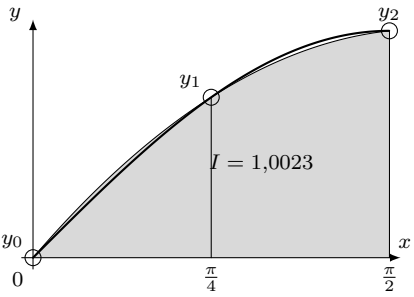
Wartość całki obliczona metodą trapezów z krokiem $2h$ wynosi

$$T(2h) = \frac{\pi}{2}\left(\frac{0+1}{2} + 0\right) = \frac{\pi}{4} = 0{,}78540$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



i	x_i	y_i
0	0	0,0
1	$\frac{\pi}{4}$	0,70711
2	$\frac{\pi}{2}$	1,0



Rysunek 4.6. Wartości funkcji podcałkowej, konstrukcja trapezów i parabola interpolująca funkcję sinus w metodzie Simpsona

Wartość całki obliczona metodą trapezów z krokiem h wyniesie

$$T(h) = \frac{\pi}{4}\left(\frac{0+1}{2} + 0,70711\right) = 0,94806$$

Wartość obliczona za pomocą ekstrapolacji Richardsona wynosi

$$I \approx T(h) + \frac{1}{3}(T(h) - T(2h)) = 0,94806 + \frac{1}{3}(0,94806 - 0,78540) = 1,00233$$

Metoda Simpsona. Przyjęty zostanie krok $h = \pi/4$, wartości funkcji podcałkowej zostaną wzięte z tabeli na rysunku 4.6, zatem wartość całki wyniesie

$$I \approx \frac{h}{3}\left(y_0 + 4y_1 + y_2\right) = \frac{\pi}{12}(0 + 4 \cdot 0,70711 + 1) = 1,00233$$

Błędy obliczeń Uzyskane wyniki należy porównać z dokładną wartością obliczanej całki, która wynosi 1. Miarę błędu stanowi różnica między wartością dokładną a uzyskaną za pomocy metody numerycznej. Takie porównanie zamieszczono poniżej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Metoda	Wartość	Błąd [%]
Trapezów $h = \pi/2$	0,78540	21,5
Trapezów $h = \pi/4$	0,94806	5,2
Ekstrapolacja Richardsona	1,0023	0,23
Metoda Simpsona $h = \pi/4$	1,0023	0,23

Powyższe obliczenia potwierdzają przewidywania teoretyczne związane z wielkością błędu w metodzie trapezów. Widać też znaczne korzyści z zastosowania ekstrapolacji Richardsona. Można też eksperymentalnie zweryfikować twierdzenia na temat równoważności metody trapezów z ekstrapolacją Richardsona i metody Simpsona.

Niniejszy przykład ilustruje korzyści ze stosowania metod wyższych rzędów. Z rysunku 4.6 bardzo dobrze widać, dlaczego przy tym samym kroku znacznie lepiej jest użyć paraboli do interpolacji funkcji podcałkowej niż wielomianów pierwszego stopnia. To spostrzeżenie ma duże znaczenie przy rozwiązywaniu równań różniczkowych metodami elementów skończonych i brzegowych. Wyniki pozwalają potwierdzić fakt, iż stosowanie ekstrapolacji Richardsona do metody trapezów, powoduje w rzeczywistości zwiększenie rzędu metody.

4.6. Kwadratury Newtona–Cotesa

Tradycyjnie kwadraturą nazywa się formułę pozwalającą wyznaczyć wartość całki oznaczonej w następujący sposób

$$\int_a^b f(x)\mathrm{d}x \approx \sum_i A_i f(x_i) \tag{4.24}$$

czyli wartość całki jest w przybliżeniu równa kombinacji liniowej wartości funkcji podcałkowej $f(x)$ w wybranych punktach x_i i wag A_i. Punkty x_i są nazywane węzłami kwadratury.

Podstawowym problemem przy konstruowaniu kwadratur zatem staje się wyznaczenie współczynników stojących przy wartościach funkcji dla zadanych węzłów x_i. Innym problemem jest określenie punktów, w których wyznaczane są potrzebne wartości funkcji. W dalszej części rozdziału zostaną omówione dwie podstawowe rodziny kwadratur.

Idea kwadratur Newtona–Cotesa polega na aproksymacji (interpolacji) funkcji podcałkowej w n równoodległych punktach za pomocą wielomianu. Następnie przeprowadza się całkowanie otrzymanego wielomianu. Należy brać pod uwagę fakt, że istnieje zjawisko Rungego. Polega ono na pojawianiu się oscylacji wielomianu interpolującego przy interpolacji z węzłami równoodległymi. Przez to gwałtownie narastają

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

błędy interpolacji. Czyli im wyższy stopień wielomianu interpolacyjnego, tym większe mogą być jego oscylacje przy węzłach równoodległych [23].

Jeżeli wprowadzi się nieujemną i całkowalną funkcję wagową $w(x)$, to wykorzystując wzór interpolacyjny Lagrange'a (2.13), można przeprowadzić interpolację funkcji podcałkowej w $m+1$ równoodległych węzłach. Finalnie uzyskuje się następującą postać wzoru (4.24)

$$\int_a^b f(x)w(x)\mathrm{d}x \approx \int_a^b L(x)w(x)\mathrm{d}x = \int_a^b \sum_{i=0}^{m} f(x_i)l_i(x)w(x)\mathrm{d}x = \\ = \sum_{i=0}^{m} f(x_i)\int_a^b l_i(x)w(x)\mathrm{d}x = \sum_{i=0}^{m} A_i f(x_i) \quad (4.25)$$

Jak widać z wzoru (4.25), współczynniki stojące przy wartościach funkcji wyznacza się zgodnie z ogólną zależnością

$$A_i = \int_a^b l_i(x)w(x)\mathrm{d}x$$

gdzie $l_i(x)$ jest wielomianem Lagrange'a. Wzór (4.25) jest dokładny dla wszystkich wielomianów stopnia m, a w niektórych przypadkach dla wielomianów stopnia wyższego niż m.

Zestawienie wybranych kwadratur Newtona–Cotesa przedstawia tabela 4.1.

Tabela 4.1. Współczynniki wybranych kwadratur Newtona–Cotesa i oszacowania błędu lokalnego

N	w	Współczynniki A_i/w							Błąd
1	$\frac{1}{2}h$	1	1						$h^3\frac{1}{12}f^{(2)}(\xi)$
2	$\frac{1}{3}h$	1	4	1					$h^5\frac{1}{90}f^{(4)}(\xi)$
3	$\frac{3}{8}h$	1	3	3	1				$h^5\frac{3}{80}f^{(4)}(\xi)$
4	$\frac{4}{90}h$	7	32	12	32	7			$h^7\frac{8}{945}f^{(6)}(\xi)$
5	$\frac{5}{288}h$	19	75	50	50	75	19		$h^7\frac{275}{12096}f^{(6)}(\xi)$
6	$\frac{6}{840}h$	41	216	27	272	27	216	41	$h^9\frac{9}{1400}f^{(8)}(\xi)$

Oprócz współczynników poszczególnych kwadratur zamieszczono w niej też oszacowanie błędu dla każdej z nich. Szacowanie wielkości błędu można przeprowadzić,

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdy jest znana jawna postać funkcji podcałkowej. W większości praktycznych przypadków całkuje się funkcje w postaci dyskretnej. Trudno jest wtedy szacować błąd w oparciu o podane formuły. W takim przypadku zdecydowanie korzystniej jest wykorzystać metodę Romberga i na jej podstawie oszacować błąd wyznaczenia całki. W sytuacji, gdy znana jest jawna postać funkcji podcałkowej, można wykorzystać znacznie dokładniejsze kwadratury Gaussa.

4.7. Kwadratury Gaussa

W odróżnieniu od kwadratur Newtona–Cotesa, przy konstrukcji kwadratur Gaussa zakłada się stopień i rodzaj wielomianu interpolującego funkcję podcałkową, natomiast poszukuje się punktów, które będą węzłami interpolacji. Poszukiwania prowadzi się w taki sposób, by zminimalizować wielkość błędu popełnianego przy całkowaniu.

Zazwyczaj minimalizację błędu popełnianego przy całkowaniu realizuje się przez odpowiedni dobór wielomianów interpolujących funkcję podcałkową. Następnie wybiera się wartości funkcji podcałkowej w miejscach zerowych wielomianów $\varphi_{m+1}(x)$ stopnia $m+1$ z rodziny wielomianów ortogonalnych związanej z funkcją wagową $w(x)$. W takim przypadku wzór (4.24) daje wartości dokładne dla wielomianów stopnia $2m+1$ [5].

Właśnie takie wielomiany wykorzystuje się w kwadraturach Gaussa. Zatem metoda pozwala osiągnąć większe dokładności od kwadratur Newtona–Cotesa, niemniej słabo się nadaje do całkowania funkcji w postaci dyskretnej, gdyż błędy popełniane przy interpolacji wartości funkcji danej w postaci tablicy znacznie pogarszają dokładność wyniku.

Aby zilustrować sposób otrzymywania kwadratur Gaussa, zostanie wyznaczona formuła w postaci kombinacji liniowej wartości funkcji $f(x_i)$ i wag w_i, pozwalająca na obliczenie wartości całki z funkcji $f(x)$.

$$I = \int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)$$

Punkty x_i nazywane są węzłami kwadratury lub punktami Gaussa. Rozmieszczenie węzłów dobiera się w taki sposób, by uzyskać wartość dokładną całki dla wielomianów możliwie wysokiego stopnia.

Kwadratura jednopunktowa. Najprostszy przypadek dotyczy sytuacji, gdy wartość całki wyznacza się na podstawie jednej wartości funkcji. Położenie węzła kwadratury i wagi ustala się tak, by dla wielomianów stopnia zerowego i pierwszego kwadratura dawała wartości dokładne. Kwadratura ma ogólną postać

$$I = \int_{-1}^{1} f(x)\mathrm{d}x \approx w_1 f(x_1) a$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Aby wyznaczyć dwie nieznane wartości w_1 i x_1, można skorzystać z następujących warunków:

— wielomian stopnia zerowego $f(x) = 1$, czyli $\int_{-1}^{1} 1\mathrm{d}x = 2$, czyli $w_1 = 2$,
— wielomian stopnia pierwszego $f(x) = x$, czyli $\int_{-1}^{1} x\mathrm{d}x = 0$, czyli $x_1 = 0$.

Zatem finalna postać kwadratury jest następująca

$$I = \int_{-1}^{1} f(x)\mathrm{d}x \approx 2f(0)$$

Kwadratura jednopunktowa daje wyniki dokładne przy całkowaniu wielomianów pierwszego stopnia, natomiast dla innych da wyniki przybliżone.

Kwadratura dwupunktowa wykorzystuje dwie wartości funkcji do wyznaczenia wartości całki, czyli

$$I = \int_{-1}^{1} f(x)\mathrm{d}x \approx w_1 f(x_1) + w_2 f(x_2)$$

Do wyznaczenia są cztery niewiadome: w_1, w_2, x_1 oraz x_2. Zakładając, że kwadratura ma dawać wyniki dokładne dla wielomianów kolejnych stopni, można napisać zależności pozwalające na wyznaczenie tych stałych:

— wielomian stopnia zerowego $f(x) = 1$, czyli $\int_{-1}^{1} 1\mathrm{d}x = 2$, czyli $w_1 + w_2 = 2$,
— wielomian stopnia pierwszego $f(x) = x$, czyli $\int_{-1}^{1} x\mathrm{d}x = 0$, czyli $w_1 x_1 + w_2 x_2 = 0$,
— wielomian stopnia drugiego $f(x) = x^2$, czyli $\int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3}$, czyli $w_1 x_1^2 + w_2 x_2^2 = \frac{2}{3}$,
— wielomian stopnia trzeciego $f(x) = x^3$, czyli $\int_{-1}^{1} x^3\mathrm{d}x = 0$, czyli $w_1 x_1^3 + w_2 x_2^3 = 0$.

Po rozwiązaniu tego układu równań, uzyskuje się następujące wartości poszukiwanych niewiadomych

$$w_1 = w_1 = 1, \qquad x_1 = -\frac{1}{\sqrt{3}}, \qquad x_2 = \frac{1}{\sqrt{3}}$$

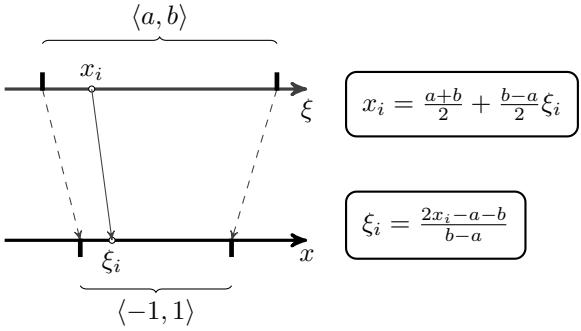
Czyli dwupunktowa kwadratura Gaussa ma postać

$$I = \int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right)$$

W podobny sposób można wyznaczyć kolejne kwadratury. Warto zauważyć, że kwadratury w tej postaci są dokładne dla wielomianów stopnia $2n-1$, gdzie n jest liczbą węzłów kwadratury.

Jak można było zauważyć, kwadratury Gaussa są określone dla całek w granicach od -1 do 1, aby je wykorzystać praktycznie, należy odwzorować przedział $\langle a, b\rangle$ na przedział $\langle -1, 1\rangle$, co zostało schematycznie przedstawione na rysunku 4.7.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.7. Schemat odwzorowania przedziału $\langle a, b\rangle$ na przedział $\langle -1, 1\rangle$

Zatem obliczenia wartości całki wykonuje się zgodnie z następującą zależnością

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{a+b}{2}+\frac{b-a}{2}\xi\right)\mathrm{d}\xi \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f\left(\frac{a+b}{2}+\frac{b-a}{2}\xi_i\right) \tag{4.26}$$

Jak zostało pokazane, współrzędne węzłów i wartości wag można stosunkowo łatwo wyznaczać dla prostych przypadków. Aby zaoszczędzić pracy czytelnikowi, w tabeli 4.2 zestawiono współrzędne węzłów i wartości wag dla dziewięciu początkowych kwadratur Gaussa. Współczynniki zostały podane ze znacznie większą precyzją niż wymagana przez 64-bitowe zmienne zmiennoprzecinkowe, aby w razie potrzeby wykorzystać je w kodach o większej długości zmiennych. Pierwsza kolumna zawiera liczbę punktów używanych w kwadraturze. W drugiej kolumnie znajdują się współrzędne węzłów, natomiast ostatnia kolumna zawiera wagi odpowiadające węzłom.

Tabela 4.2. Kwadratury Gaussa–Legendre'a

	Węzły	Wagi
2	±0,577350269189625764509148 8	1,000000000000000000000000
3	0	0,888888888888888888888889
	±0,774596669241483377035853 1	0,555555555555555555555556
4	±0,339981043584856264802665 8	0,652145154862546142626936 1
	±0,861136311594052575223946 5	0,347854845137453857373063 9
5	0	0,568888888888888888888889

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 4.2. cd.

	±0,53846931010568309103631 44	0,47862867049936646804129 15
	±0,90617984593866399279762 69	0,23692688505618908751426 40
6	±0,23861918608319690863050 17	0,46791393457269104738987 03
	±0,66120938646626451366139 96	0,36076157304813860756983 35
	±0,93246951420315202781230 16	0,17132449237917034504029 61
7	0	0,41795918367346938775510 20
	±0,40584515137739716690660 64	0,38183005050511894495036 98
	±0,74153118559939443986386 48	0,27970539148927666790146 78
	±0,94910791234275852452618 97	0,12948496616886969327061 14
8	±0,18343464249564980493947 61	0,36268378337836198296515 04
	±0,52553240991632898581773 90	0,31370664587788728733796 22
	±0,79666647741362673959155 39	0,22238103445337447054435 60
	±0,96028985649753623168356 09	0,10122853629037625915253 14
9	0	0,33023935500125976316452 51
	±0,32425342340380892903853 80	0,31234707704000284006863 04
	±0,61337143270059039730870 20	0,26061069640293546231874 29
	±0,83603110732663579429942 98	0,18064816069485740405847 20
	±0,96816023950762608983557 62	0,08127438836157441197189 22
10	±0,14887433898163121088482 60	0,29552422471475287017389 30
	±0,43339539412924719079926 59	0,26926671930999635509122 69
	±0,67940956829902440623432 74	0,21908636251598204399553 49
	±0,86506336668898451073209 67	0,14945134915058059314577 63
	±0,97390652851717172007796 40	0,06667134430868813759356 88

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

4.7.1. Kwadratury Gaussa–Kronroda

Przy obliczeniach numerycznych zawsze występuje problem określenia dokładności obliczeń. Z tego powodu opracowuje się metody pozwalające nie tylko rozwiązywać zagadnienia, ale też oszacowywać dokładność tych rozwiązań. Dlatego też opracowane zostały kwadratury Gaussa–Kronroda. Kwadratury Gaussa–Kronroda zostały opracowane w połowie lat 60. ubiegłego wieku przez rosyjskiego informatyka Aleksandra Siemionowicza Kronroda. Wykorzystuje dwa zestawy węzłów. Węzły kwadratury Gaussa oraz węzły Kronroda. Przy czym węzłów Kronroda jest znacznie więcej i zawierają one węzły Gaussa. Węzły Kronroda wyznaczane są jako miejsca zerowe wielomianów Stieltjesa. Zawsze występują w parach. Pary zapisuje się, używając litery G z liczbą węzłów na oznaczenie kwadratur Gaussa biorących udział w obliczeniach oraz litery K z liczbą węzłów na oznaczenie węzłów Kronroda. Najpopularniejsze pary to $(G7, K15)$, $(G10, K21)$, $(G15, K31)$, $(G20, K41)$, $(G25, K51)$, $(G30, K61)$. Przykładowe węzły i wagi dla dwóch kwadratur Gaussa–Kronroda $(G7, K15)$ oraz $(G10, K21)$ przedstawiono w tabelach 4.3 oraz 4.4.

Tabela 4.3. Kwadratura Gaussa–Kronroda $G7, K15$

Węzły Gaussa		Wagi
$0{,}00000000000000000000$	*	$4{,}17959183673469387755 \cdot 10^{-1}$
$\pm 4{,}05845151377397166907 \cdot 10^{-1}$	*	$3{,}81830050505118944950 \cdot 10^{-1}$
$\pm 7{,}41531185599394439864 \cdot 10^{-1}$	*	$2{,}79705391489276667901 \cdot 10^{-1}$
$\pm 9{,}49107912342758524526 \cdot 10^{-1}$	*	$1{,}29484966168869693270 \cdot 10^{-1}$
Węzły Kronroda		**Wagi**
$0{,}00000000000000000000$	*	$2{,}09482141084727828013 \cdot 10^{-1}$
$\pm 2{,}07784955007898467601 \cdot 10^{-1}$		$2{,}04432940075298892414 \cdot 10^{-1}$
$\pm 4{,}05845151377397166907 \cdot 10^{-1}$	*	$1{,}90350578064785409913 \cdot 10^{-1}$
$\pm 5{,}86087235467691130294 \cdot 10^{-1}$		$1{,}69004726639267902827 \cdot 10^{-1}$
$\pm 7{,}41531185599394439864 \cdot 10^{-1}$	*	$1{,}40653259715525918745 \cdot 10^{-1}$
$\pm 8{,}64864423359769072790 \cdot 10^{-1}$		$1{,}04790010322250183839 \cdot 10^{-1}$
$\pm 9{,}49107912342758524526 \cdot 10^{-1}$	*	$6{,}30920926299785532907 \cdot 10^{-2}$
$\pm 9{,}91455371120812639207 \cdot 10^{-1}$		$2{,}29353220105292249637 \cdot 10^{-2}$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

W środkowej kolumnie gwiazdki są w tych wierszach, w których znajdują się węzły, które są identyczne w kwadraturach Gausa oraz Kronroda. Ponieważ węzły obu kwadratur są identyczne, więc wystarczy raz wyznaczyć wartości w węzłach Kronroda, by móc obliczyć wartość kwadratury Kronroda i wartość odpowiedniej kwadratury Gaussa. Ma to niebagatelne znacznie, ponieważ umożliwia oszacowanie błędu całkowania.

Tabela 4.4. Kwadratura Gaussa–Kronroda $G10, K21$

Węzły Gaussa		Wagi
$\pm 1{,}48874338981631210885 \cdot 10^{-1}$	*	$2{,}95524224714752870173 \cdot 10^{-1}$
$\pm 4{,}33395394129247190799 \cdot 10^{-1}$	*	$2{,}69266719309996355091 \cdot 10^{-1}$
$\pm 6{,}79409568299024406234 \cdot 10^{-1}$	*	$2{,}19086362515982043996 \cdot 10^{-1}$
$\pm 8{,}65063366688984510732 \cdot 10^{-1}$	*	$1{,}49451349150580593146 \cdot 10^{-1}$
$\pm 9{,}73906528517171720079 \cdot 10^{-1}$	*	$6{,}66713443086881375936 \cdot 10^{-2}$
Węzły Kronroda		**Wagi**
$0{,}00000000000000000000$		$1{,}49455554002916905665 \cdot 10^{-1}$
$\pm 1{,}48874338981631210885 \cdot 10^{-1}$	*	$1{,}47739104901338491375 \cdot 10^{-1}$
$\pm 2{,}94392862701460198131 \cdot 10^{-1}$		$1{,}42775938577060080797 \cdot 10^{-1}$
$\pm 4{,}33395394129247190799 \cdot 10^{-1}$	*	$1{,}34709217311473325928 \cdot 10^{-1}$
$\pm 5{,}62757134668604683339 \cdot 10^{-1}$		$1{,}23419176262065851078 \cdot 10^{-1}$
$\pm 6{,}79409568299024406234 \cdot 10^{-1}$	*	$1{,}09387158802297641899 \cdot 10^{-1}$
$\pm 7{,}80817726586416897064 \cdot 10^{-1}$		$9{,}31254545836976055351 \cdot 10^{-2}$
$\pm 8{,}65063366688984510732 \cdot 10^{-1}$	*	$7{,}50396748109199527670 \cdot 10^{-2}$
$\pm 9{,}30157491355708226001 \cdot 10^{-1}$		$5{,}47558965743519960314 \cdot 10^{-2}$
$\pm 9{,}73906528517171720078 \cdot 10^{-1}$	*	$3{,}25581623079647274788 \cdot 10^{-2}$
$\pm 9{,}95657163025808080736 \cdot 10^{-1}$		$1{,}16946388673718742781 \cdot 10^{-2}$

Jak już wspomniano, zasadniczą zaletą kwadratur Gaussa–Kronroda jest możliwość oszacowania błędu wyznaczenia wartości całki. Na przykład dla pary kwadratur

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$(G7, K15)$ oszacowanie błędu można uzyskać na podstawie wyrażenia

$$\Delta = (200 \cdot |G7 - K15|)^{1,5}$$

gdzie $G7$ jest wartością całki wyznaczoną za pomocą siedmiopunktowej kwadratury Gaussa, zaś $K15$ jest wartością tej samej całki wyznaczoną za pomocą piętnastopunktowej kwadratury Kronroda. Podobne wyrażenia opisujące wielkość błędu zostały opracowane dla innych par kwadratur i pozwalają oszacować jakość rozwiązania. Przy niezadowalającej wielkości błędu należy użyć kwadratur wyższego rzędu, czyli zawierających więcej węzłów, co pozwoli zwiększyć dokładność rozwiązania.

4.7.2. Kilka słów podsumowania

Czytelnik koniecznie powinien zdawać sobie sprawę z następujących zależności:

1. przy $N + 1$ węzłach rząd kwadratury Gaussa wynosi $2N + 2$ zaś rząd kwadratury Newtona–Cotesa $N + 2$;
2. przy tej samej liczbie węzłów kwadratury Gaussa są znacznie dokładniejsze niż kwadratury Newtona–Cotesa;
3. aby unikać obniżającej dokładność interpolacji, stablicowane funkcje czy sygnały próbkowane lepiej całkować kwadraturami Newtona–Cotesa;

Oprócz tego warto też pamiętać o możliwości poprawiania dokładności metodą ekstrapolacji Richardsona.

4.8. Metody Monte Carlo

Podobnie jak metody wymienione w poprzednich rozdziałach metody Monte Carlo zostaną omówione na przykładzie całki pojedynczej. W rzeczywistości metody Monte Carlo stanowią obszerny dział z pogranicza metod statystycznych i numerycznych. Ponieważ celem niniejszego rozdziału jest przybliżenie tych metod, zostanie przedstawiony jedynie zarys tego złożonego zagadnienia.

Przy wyznaczeniu wartości całki oznaczonej można wykorzystać jej interpretację geometryczną w jeszcze inny niż do tej pory sposób. Niech fragment płaszczyzny będzie traktowany jako zbiór punktów. Pole powierzchni w pewien sposób charakteryzuje liczbę tych punktów. Położenie każdego z punktów jest opisane współrzędnymi (x, y). Przy wyznaczaniu wartości całki (4.1), prostokątny fragment płaszczyzny zostanie ograniczony prostymi

$$x = a, \quad x = b, \quad y = 0, \quad y = c$$

gdzie c jest lokalnym maksimum funkcji $f(x)$ w przedziale $\langle a, b\rangle$.

Sama metoda Monte Carlo polega na wielokrotnym losowaniu dwóch wartości, które będą współrzędnymi punktu. Współrzędne punku $A(x_1, x_2)$ stanowią realizacje zmiennych losowych X_1 oraz X_2 o rozkładach równomiernych, odpowiednio $U(a, b)$ oraz $U(0, c)$. Po wylosowaniu współrzędnych punktu, dla każdego z nich następuje sprawdzenie warunku

$$x_2 < f(x_1)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Spełnienie warunku oznacza, że punkt leży pod funkcją i należy do obszaru całki. Po zliczeniu punktów spełniających warunek i oznaczeniu ich liczby przez N_1, zaś liczby wszystkich wylosowanych punktów przez N, można zapisać wyrażenie na wartość całki oznaczonej jako

$$\int_a^b f(x)\mathrm{d}x \approx (b-a)\, c\, \frac{N_1}{N} \tag{4.27}$$

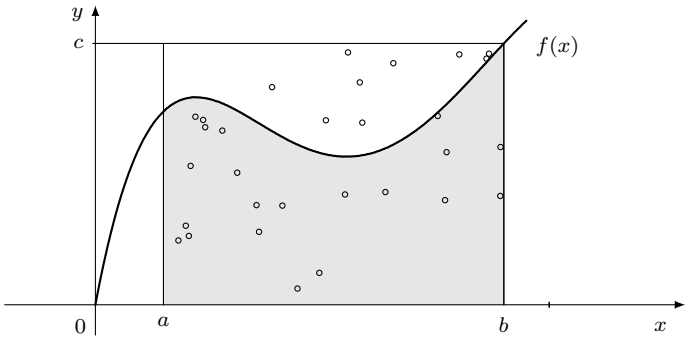
Wynik obliczeń tą metodą jest zmienną losową, zaś dokładność metody silnie zależy od jakości generatora liczb losowych oraz od liczby zrealizowanych losowań.

4.8.1. Przykład obliczeniowy

Sformułowanie problemu. Posługując się metodą Monte Carlo, wyznaczyć wartość całki oznaczonej

$$\int_{0{,}15}^{0{,}9} (-4x^6 + 18x^5 - 42x^4 + 53x^3 - 32x^2 + 8x)\mathrm{d}x$$

Rysunek 4.8 przedstawia funkcję podcałkową z granicami. Pole powierzchni zaciemnionego obszaru to wartość poszukiwanej całki oznaczonej.



Rysunek 4.8. Przykład metody Monte Carlo dla całek oznaczonych

Rozwiązanie. Na rysunku 4.8 zaznaczono maksymalną wartość $c = f(b) = 0{,}8638$ funkcji podcałkowej w analizowanym przedziale. Obliczenia poprowadzono w następujący sposób:

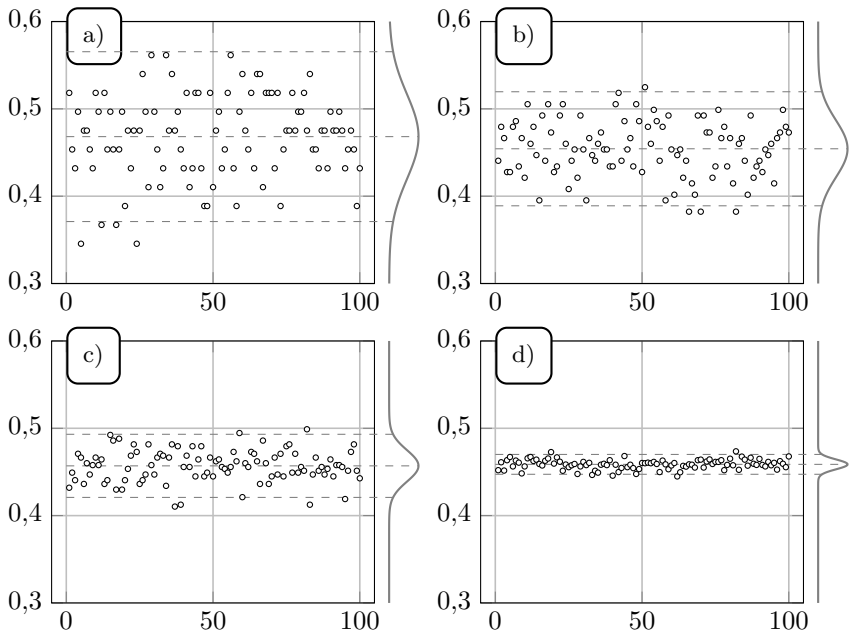
— Przyjęto $N = 30$ punktów.
— Wygenerowano N par liczb pseudolosowych o rozkładzie równomiernym $x_1 = U(0{,}15; 0{,}9)$ oraz $x_2 = U(0; 0{,}8638)$.
— Określono liczbę punktów $A_i(x_1, x_2)$ znajdujących się pod funkcją podcałkową $N_1 = 21$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zatem, zgodnie z zależnością (4.27), wynik obliczeń całki metodą Monte Carlo wyniesie

$$I_p = 0{,}8638(0{,}9 - 0{,}15)\frac{21}{30} = 0{,}45353$$

Dokładna wartość całki $I_d = 0{,}45812$. Błąd względny popełniony przy wyznaczeniu wartości całki wynosi około 1%.



Rysunek 4.9. Wartości całki oznaczonej obliczonej 100 razy metodą Monte Carlo, przy losowaniu a) 30, b) 100, c) 300 i d) 3000 punktów z zaznaczeniem obszaru 2σ

Uważny czytelnik przekona się, że duża dokładność wyniku jest raczej przypadkiem. Jest to oczywiście naturalna cecha metody Monte Carlo. Jak widać, liczba wylosowanych punktów jest raczej daleka od wystarczającej. Kolejne obliczenia dla tej samej liczby punktów dają wyniki przedstawione na rysunku 4.9a. W jednej z serii wartość średnia ze 100 realizacji obliczeń wynosi $\bar{I} = \mu = 0{,}46195$, co daje błąd obliczeń około 0,8%. Odchylenie standardowe średniej wynosi $\sigma = 0{,}051354$. Na poszczególnych wykresach rysunku 4.9 kreskowanymi liniami poziomymi zaznaczono przedział $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$, w którym powinno się mieścić 95% wyników. Na wykresie b) przedstawiono wyniki stu obliczeń wartości całki przy losowaniu 100 punktów. Wykres c) przedstawia wyniki dla 300 punktów, zaś wykres d) wyliczone wartości całki przy losowaniu 3000 punktów przy każdorazowym wyznaczeniu wartości całki.

Skala na osi pionowej (wartość całki) jest identyczna na każdym z rysunków. Widać jednoznacznie, iż wraz ze wzrostem liczby losowanych punktów, rośnie dokładność wyznaczenia całki oznaczonej metodą Monte Carlo. Dobrze też widać, iż przy małej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

liczbie losowanych punktów pojawia się zjawisko kwantyzacji, czyli wartość całki nie przyjmuje dowolnych wartości, a jedynie te z dozwolonego zbioru. Wynika to ze sposobu obliczania wartości całki – iloraz P_1/P_2 dla P_1 i P_2 naturalnych może przyjąć wartości z mało licznego zbioru.

Na poszczególnych wykresach naszkicowano kształt funkcji gęstości prawdopodobieństwa rozkładu normalnego $N(\mu, \sigma)$ o parametrach μ i σ wyznaczonych dla obliczeń z różnymi liczbami losowanych punktów.

Zwiększanie liczby wylosowanych punktów „próbnych", służących do wyznaczenia całki, poprawia dokładność estymacji wartości całki oznaczonej. Można też łatwo określić wpływ liczby wylosowanych punktów na dokładność wyniku.

Opisana metoda to tylko prosty przykład metod statystycznych stosowanych do rozwiązywania problemów natury deterministycznej. Metody Monte Carlo to liczna rodzina metod stosowanych w różnych działach fizyki i znajdujących zastosowanie w technice.

4.9. Przykłady praktyczne

4.9.1. Moc sprężarki

Sformułowanie problemu. Na rysunku 4.10 przedstawiono wykres indykatorowy jednostopniowej sprężarki rzeczywistej. Na podstawie zamieszczonego wykresu wyznaczyć pracę wykonaną w trakcie jednego cyklu pracy. Znając prędkość obrotową $n = 130$ obr/min, wyznaczyć zapotrzebowanie na moc sprężarki. Wyznaczyć średni wykładnik politropy przy sprężaniu.

Rozwiązanie. W układzie $p - V$ pole powierzchni zamknięte krzywą $p(V)$ wyznacza pracę przekazaną do gazu w procesie jego sprężania. Pole to zostało zaciemnione na rysunku 4.10. Sama krzywa $p(V)$ została podzielona na dwie funkcje: $p_d(V)$ oraz $p_g(V)$. Podział został dokonany w odczytanych z wykresu punktach, związanych z granicznymi wartościami objętości cylindra sprężarki.

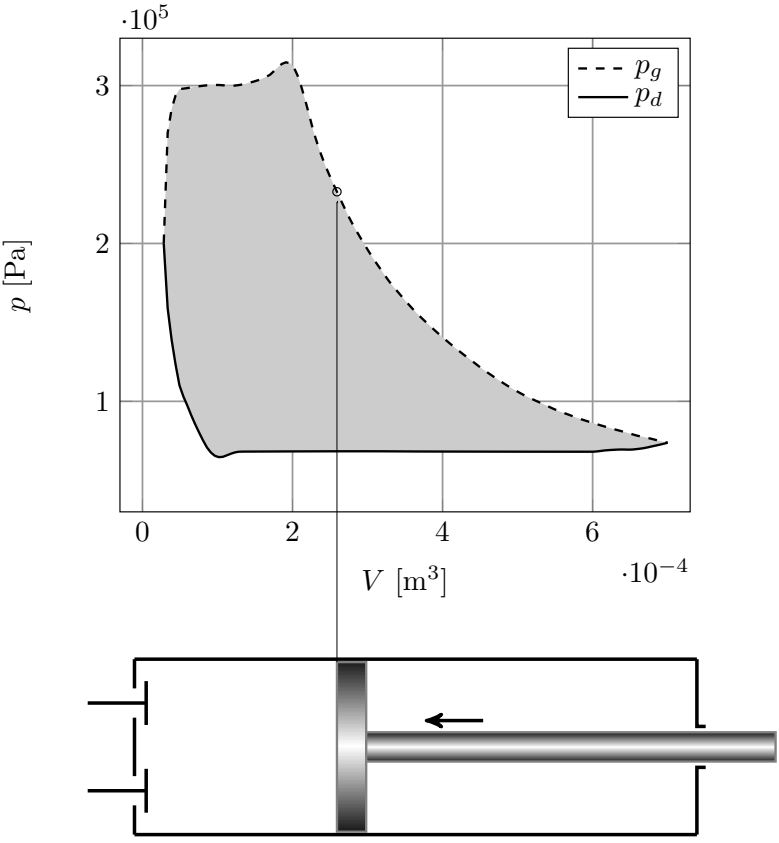
Aby wyznaczyć wielkość ciemniejszego pola, należy wyznaczyć wartość następującej całki

$$L = \int_{V_{min}}^{V_{max}} p_g - p_d \, \mathrm{d}V \tag{4.28}$$

gdzie V_{min} jest objętością cylindra przy lewym skrajnym położeniu, zaś V_{max} jest objętością cylindra w prawym skrajnym położeniu. Funkcja $p_g(V)$ została na rysunku 4.10 narysowana linią przerywaną, natomiast funkcja $p_d(V)$ – linią ciągłą.

Aby możliwe było wyznaczenie wartości takiej całki, konieczne było odczytanie współrzędnych punktów leżących na krzywej wykresu indykatorowego. Cały zakres ruchu tłoka został podzielony na 32 przedziały, na granicach których odczytano wartości ciśnień na dolnej – p_{di} i górnej – p_{gi} krzywej. Odczytane wartości zostały przedstawione w tabeli 4.5.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.10. Wykres indykatorowy sprężarki

W pierwszym kroku zostały wyznaczone wartości funkcji podcałkowej, czyli obliczono wartości różnic $p_{gi} - p_{di}$, które przedstawiono na rysunku 4.11. Było to możliwe dzięki temu, że – jak widać z tabeli 4.5 – wartości obu ciśnień były odczytywane z wykresu indykatorowego dla tych samych przemieszczeń tłoka. Jeżeli ciśnienia p_g oraz p_d nie są odczytywane w tych samych położeniach tłoka, to całkę (4.28) trzeba rozbić na dwie i wartość każdej z nich wyznaczać oddzielnie, co prowadzi do następującego wyrażenia

$$L = \int_{V_{min}}^{V_{max}} p_g - p_d \,\mathrm{d}V = \int_{V_{min}}^{V_{max}} p_g \,\mathrm{d}V - \int_{V_{min}}^{V_{max}} p_d \,\mathrm{d}V$$

Jeżeli odczyt wartości ciśnień odbywa się ze stałym krokiem, różnym dla górnej i dolnej krzywej, do każdej z tych całek można użyć metody Romberga. Jeżeli krok, z którym odbywa się odczyt, nie jest stały, pozostaje wykorzystanie metody trapezów lub opracowanie własnej kwadratury wykorzystującej interpolację wielomianem niskiego stopnia przy węzłach nierównoodległych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

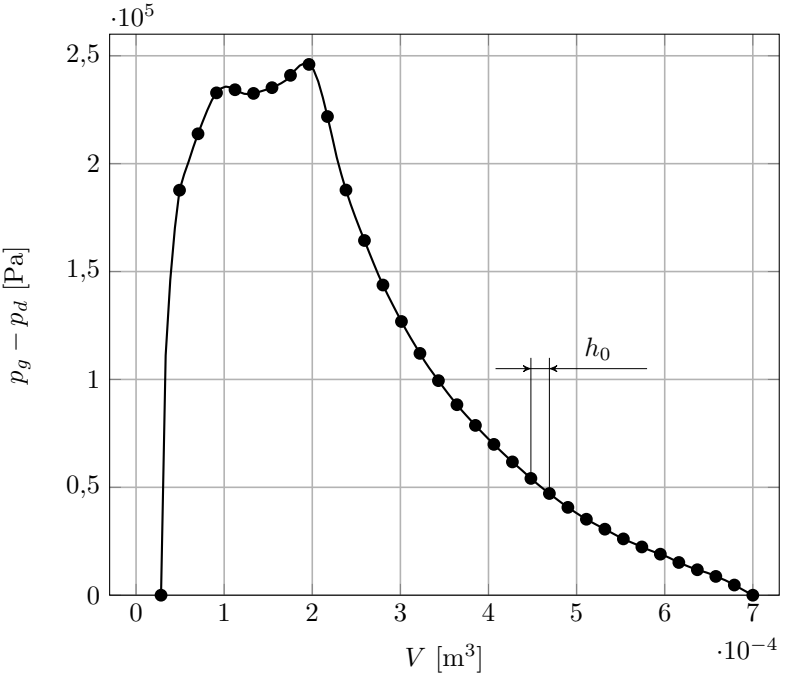
Tabela 4.5. Wartości odczytane z wykresu indykatorowego

i	V_i [cm^3]	p_{di} [kPa]	p_{gi} [kPa]	i	V_i [cm^3]	p_{di} [kPa]	p_{gi} [kPa]
1	28,343	200,147	200,147	18	385,161	68,211	146,905
2	49,332	110,083	297,786	19	406,150	68,191	138,087
3	70,321	85,416	299,256	20	427,139	68,172	129,948
4	91,311	67,525	300,336	21	448,129	68,154	122,277
5	112,300	65,764	300,023	22	469,118	68,137	115,269
6	133,289	68,156	300,714	23	490,107	68,123	108,830
7	154,279	68,199	303,417	24	511,096	68,110	103,318
8	175,268	68,239	309,171	25	532,086	68,100	98,695
9	196,257	68,276	314,272	26	553,075	68,092	94,192
10	217,246	68,309	290,148	27	574,064	68,087	90,454
11	238,236	68,337	256,110	28	595,054	68,085	87,101
12	259,225	68,358	232,749	29	616,043	68,895	84,061
13	280,214	68,371	212,124	30	637,032	69,481	81,283
14	301,204	68,375	195,228	31	658,021	69,717	78,428
15	322,193	68,348	180,419	32	679,011	71,468	76,186
16	343,182	68,295	167,740	33	700,000	73,797	73,797
17	364,171	68,242	156,532	–	–	–	–

Na rysunku 4.11 została przedstawiona różnica występująca pod całką w równaniu (4.28). Zostały też zaznaczone punkty użyte do całkowania. Ta różnica zostanie scałkowana za pomocą metody Romberga. Obliczenia przedstawione w tabeli 4.6 rozpoczynają się od jednego przedziału, rozciągającego się od dolnej do górnej granicy całkowania, co oznacza, że krok całkowania wynosi $32\,h_0$. Następnie krok jest zmniejszany o połowę aż do osiągnięcia wartości h_0.

Błędami zostały nazwane poprawki wprowadzane w metodzie Romberga w postaci $\frac{1}{4^j-1}\Delta$, jak to zostało zapisane we wzorze (4.16). Wielkość pola zakreślonego przez wykres indykatorowy określa pracę wykonaną w jednym cyklu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.11. Zależność $(p_g - p_d)$ od V z zaznaczonymi wartościami i podstawowym krokiem całkowania

Z danych zawartych w tabeli 4.6 wynika, że praca wykonywana w ciągu jednego cyklu pracy wynosi $W = 74{,}067 \pm 0{,}001$. Przy czym dokładność dotyczy jedynie operacji całkowania. Podstawowym źródłem błędów jest sam wykres indykatorowy wykonywany za pomocą urządzenia mechanicznego.

Średni wykładnik politropy przy sprężaniu można wyznaczyć, obliczając wartość ilorazu pracy technicznej i pracy bezwzględnej, co można zapisać jako

$$m = \frac{L_t}{L} = \frac{\int\limits_B^A V \mathrm{d}p}{\int\limits_A^B p \,\mathrm{d}V}$$

gdzie L_t jest pracą techniczną, zaś L jest pracą absolutną. Pracę techniczną oraz pracę absolutną stanowią szare obszary na rysunku 4.12. Aby wyznaczyć ten iloraz, konieczne jest wyznaczenie wartości obu całek oznaczonych. Całkę oznaczającą pracę bezwzględną można łatwo wyznaczyć dowolną kwadraturą Newtona–Cotesa. Problemów nastręcza całka oznaczająca pracę techniczną, ponieważ to zmienna V została zdyskretyzowana ze stałym krokiem, natomiast różnice między kolejnymi dyskretnymi wartościami ciśnienia nie są stałe.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 4.6. Wyniki i oszacowane błędy obliczeń wartości całki (4.28)

h	$T(h)$	$+\frac{1}{3}\Delta$	$+\frac{1}{15}\Delta$	$+\frac{1}{63}\Delta$	$+\frac{1}{255}\Delta$	$+\frac{1}{1023}\Delta$
$32\,h_0$	0	–	–	–	–	–
$16\,h_0$	29,6505	39,534	–	–	–	–
$8\,h_0$	61,2688	71,8083	73,9599	–	–	–
$4\,h_0$	68,1886	70,4952	70,4077	70,3513	–	–
$2\,h_0$	71,4797	72,5768	72,7155	72,7522	72,7616	–
h_0	73,3313	73,9484	74,0399	74,0609	74,066	74,0673
			Błędy			
$16\,h_0$	–	9,88349	–	–	–	–
$8\,h_0$	–	10,5394	2,15162	–	–	–
$4\,h_0$	–	2,3066	0,0875374	$5{,}64 \cdot 10^{-2}$	–	–
$2\,h_0$	–	1,09704	0,138771	$3{,}66 \cdot 10^{-2}$	$9{,}42 \cdot 10^{-3}$	–
h_0	–	0,617179	0,0914451	$2{,}10 \cdot 10^{-2}$	$5{,}12 \cdot 10^{-3}$	$1{,}28 \cdot 10^{-3}$

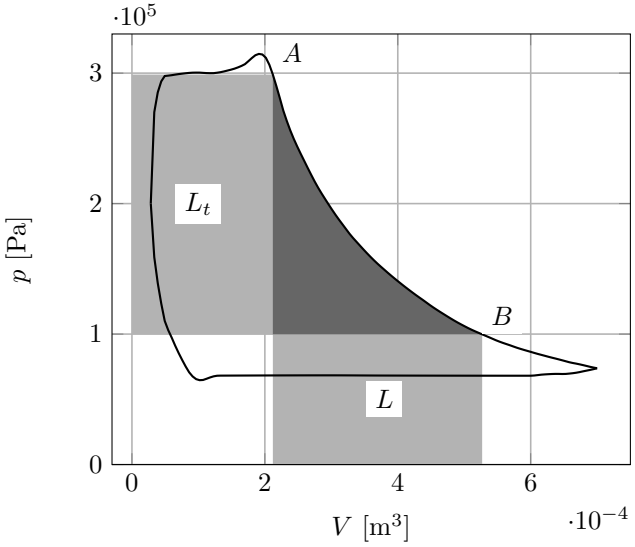
Zatem pozostaje do wyboru kilka możliwości:

— dokonać interpolacji funkcji $V(p)$, a następnie wyznaczyć jej wartości ze stałym krokiem h;
— opracować wzory, które pozwolą scałkować funkcję przy zmiennym kroku;
— posłużyć się metodą trapezów, ale obliczać pola poszczególnych trapezów i sumować je.

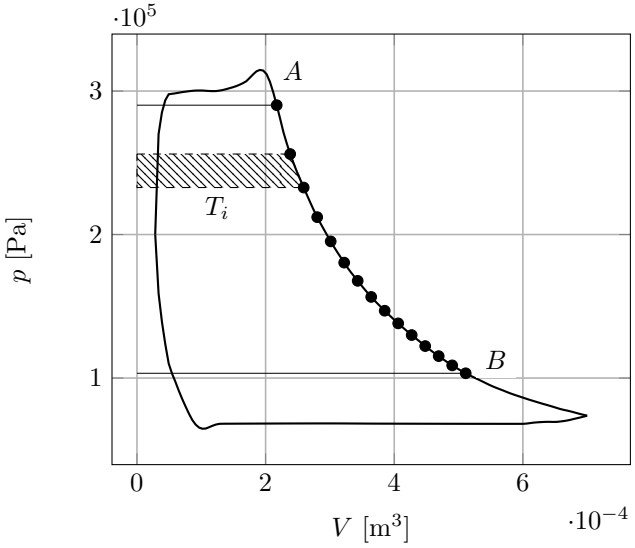
Każda z metod ma wady i zalety. Pierwsza wprowadzi dodatkowe błędy związane z interpolacją. Z dydaktycznego punktu widzenia najbardziej kusząca jest możliwość druga, sprowadzająca się do dokonania interpolacji funkcją $p(V)$ i całkowaniu funkcji odwrotnej $V(p)$. Metoda trzecia wydaje się najprostsza. Należy mieć świadomość, że metoda trapezów dla wypukłej funkcji będzie prowadzić do nadmiarowego oszacowania wartości całki.

Najszybsza będzie metoda ostatnia, to znaczy wykorzystanie metody trapezów w wersji uproszczonej, z obliczeniem pól kolejnych trapezów. Obliczenia zostaną wykonane dla danych przedstawionych w tabeli 4.5 dla kolumny oznaczonej p_g. Sprężanie zachodzi dla punktów o indeksach $i = 10$ do $i = 24$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 4.12. Praca techniczna i absolutna przy sprężaniu



Rysunek 4.13. Wyznaczanie pracy technicznej metodą trapezów

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zatem praca bezwzględna obliczona metodą trapezów będzie wynosić

$$L = h_0 \left[\frac{p_{10} + p_{24}}{2} + \sum_{i=11}^{23} p_i \right] = 49{,}513 \text{ J}$$

Wyznaczenie pracy technicznej polega na sumowaniu pól trapezów T_i, budowanych zgodnie z rysunkiem 4.13, w granicach od punktu A do punktu B. Po zrealizowaniu obliczeń dla 14 trapezów wartość wykładnika politropy wynosi

$$m = \frac{L_t}{L} = \frac{59{,}741}{49{,}513} = 1{,}2066$$

Zapotrzebowanie na moc, gdy wał sprężarki wiruje ze stałą prędkością obrotową równą n, wyznacza się za pomocą zależności

$$N = \frac{n}{60} W = \frac{130}{60} 74{,}067 = 160{,}479 \text{ W}$$

Jak widać na tym przykładzie, czasem trudno jest zastosować wprost metodę numeryczną, ale zawsze można tak ją zmodyfikować, by można było wyznaczyć poszukiwaną wartość.

4.9.2. Wymiana ciepła

Sformułowanie problemu. W zbiorniku o objętości $V = 1{,}50$ m^3 znajduje się gaz o temperaturze $t_1 = 95$ °C, którego ciśnienie bezwzględne wynosi $p_1 = 413\,325$ Pa. Zmiany molowe ciepła właściwego gazu przy zmianie temperatury opisuje zależność

$$c_n = \sqrt{342 + 0{,}96t} \left[\frac{\text{kJ}}{\text{kmol} \cdot \text{K}} \right]$$

Obliczyć ilość ciepła, jaką trzeba odebrać od gazu, by ochłodzić go do temperatury $t_2 = 15$ °C.

Rozwiązanie. Równanie pozwalające wyznaczyć ilość ciepła wymienionego przez n moli gazu ma postać

$$\mathrm{d}Q = n c_n \, \mathrm{d}t \tag{4.29}$$

Liczbę moli gazu można wyznaczyć z równania stanu gazu doskonałego jako

$$n = \frac{p_1 V}{\mu R \cdot T_1}$$

gdzie $\mu R = 8314{,}33 \text{ kJ} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}$ jest uniwersalną stałą gazową. Całkowitą ilość ciepła wymienionego przez gaz z otoczeniem można wyznaczyć, całkując równanie (4.29). Zatem ilość ciepła można określić jako

$$Q_{1-2} = n \int_{t_1}^{t_2} c_n \mathrm{d}t = \frac{p_1 V}{\mu R \cdot T_1} \int_{t_1}^{t_2} \sqrt{342 + 0{,}96t} \, \mathrm{d}t \tag{4.30}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po podstawieniu danych uzyskuje się wyrażenie

$$Q_{1-2} = 0{,}20257 \int_{t_1}^{t_2} \sqrt{342 + 0{,}96t}\,\mathrm{d}t \text{ kJ}$$

które można wygodnie całkować numerycznie na przykład metodą Romberga. Poniżej zostały zestawione wartości funkcji podcałkowej dla różnych długości kroku, które można wykorzystać do przeprowadzenia całkowania.

	t [°C]			$c_n(t)$ $\left[\frac{\text{kJ}}{\text{kmol·K}}\right]$		
i	$4h_0$	$2h_0$	$h_0 = 20$	$h = 4h_0$	$h = 2h_0$	$h_0 = 20$
1	15	15	15	18,8786	18,8786	18,8786
2	–	–	35	–	–	19,3804
3	–	55	55	–	19,8696	19,8696
4	–	–	75	–	–	20,347
5	95	95	95	20,8135	20,8135	20,8135

Zastosowanie metody Romberga do wyznaczenia wartości całki w wyrażeniu (4.30), z długościami kroku całkowania od h_0 do $4h_0$, prowadzi do tablicy jak poniżej.

h	$T(h)$	$+\frac{1}{3}\Delta$	$+\frac{1}{15}\Delta$
$4h_0$	1587,68066	–	–
$2h_0$	1588,62332	1588,93754	–
h_0	1588,85951	1588,938237	1588,938283

Błąd wyznaczenia całki można wyliczyć jako różnicę ostatnich dwóch wartości z ostatniego wiersza powyższego zestawienia i wynosi on mniej niż $4{,}66 \cdot 10^{-5}$. Zatem ilość ciepła wymienionego z otoczeniem wyniesie $Q_{1-2} = 0{,}20257 \cdot 1588{,}93828 = 321{,}8384$ kJ co jest zgodne z wynikiem wyznaczonym analitycznie.

4.10. Całkowanie w pakietach MATLAB i GNU Octave

Wyznaczanie wartości całek oznaczonych w pakietach inżynierskich MATLAB i GNU Octave można zrealizować, pisząc własny M-plik skryptowy lub funkcyjny. Jednak zdecydowanie łatwiej i bezpieczniej jest zastosować polecenia służące do całkowania numerycznego, takie jak na przykład `quad`. Wartość całki oznaczonej można

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wyznaczyć za pomocą poleceń quad, i pochodnych quadl, quadgk. Do wyznaczania wartości całek wielokrotnych stosować można polecenia dblquad czy triplequad. Ze szczegółowym opisem tych poleceń z ich możliwościami i opcjami można się zapoznać, posługując się wbudowanym systemem pomocy pakietu MATLAB. Na przykład wartość całki

$$\int_0^{\frac{\pi}{2}} \sin(\cos\theta)\mathrm{d}\theta$$

można wyznaczyć za pomocą następujących poleceń

```
1 f='sin(cos(x))'
  y=inline(f,'x')
3 calka=quad(y,0,pi/2)
```

W pierwszej linii tworzy się łańcuch tekstowy, który następnie jest przekształcany w funkcję poleceniem inline. Zaś całkowanie odbywa się z wykorzystaniem zdefiniowanej funkcji. Można też obliczenia poprowadzić tak jak poniżej

```
1 y=@(x)('sin(cos(x))')
  calka=quad(y,0,pi/2)
```

Druga metoda wykorzystuje funkcje anonimowe. Właściwe wyznaczenie wartości całki odbywa się w ostatnich liniach powyższych skryptów.

4.11. Podsumowanie

Całkowanie numeryczne pozwala uzyskiwać dokładne wartości całek oznaczonych pod warunkiem doboru właściwej metody. Opisane metody nie zawsze dadzą zadowalający wynik. Niekiedy są konieczne dodatkowe operacje na funkcji podcałkowej, zanim można przejść do całkowania numerycznego. Owe dodatkowe operacje mogą oznaczać podstawianie, całkowanie przez części, czy też analizę przebiegu funkcji podcałkowej, a głównie jej wartości na krańcach przedziału całkowania. Jest to szczególnie ważne, gdy funkcja podcałkowa ma osobliwości lub granice całki są niewłaściwe.

Opisano podstawowe metody całkowania stosowane we współczesnych systemach obliczeniowych. Można je wykorzystywać w samodzielnie tworzonych procedurach. Mimo to, czytelnik tworzący oprogramowanie obliczeniowe powinien bliżej zapoznać się z gotowymi, dojrzałymi i dogłębnie przetestowanymi procedurami z obsługą błędów i szacowaniem dokładności. Wiele z takich gotowych bibliotek jest dostępnych na wolnej licencji.

W rozdziale tym autorzy położyli duży nacisk na prostotę sformułowania, które pozwala na intuicyjne pojęcie zagadnienia całkowania numerycznego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 5

Różniczkowanie numeryczne

Omawiane w niniejszym rozdziale różniczkowanie numeryczne polega na numerycznym wyznaczaniu pochodnej funkcji w punktach, jak pokazano w równaniu (5.1). Funkcja może być dana w postaci formuły matematycznej lub w postaci tabelarycznej. W praktyce najczęściej zachodzi ten drugi przypadek, czyli funkcja jest w postaci dyskretnej.

Problem wyznaczania pochodnej na podstawie dyskretnych punktów jest trudniejszy niż problem całkowania takiej samej funkcji. Wynika to z faktu, że pochodna określa szybkość przyrostu wartości funkcji. Przy braku informacji na temat zachowania funkcji między punktami ze znanymi wartościami, konieczne są silne założenia co do jej charakteru lub rodzaju. Bez takich założeń należy się liczyć z dużymi błędami aproksymacji pochodnej. W związku z tym należy stosować małe długości kroku dla argumentów funkcji różniczkowanych. Niestety, małe długości kroku są bardzo niebezpieczne, ponieważ najczęściej powodują niewielkie różnice w wartościach funkcji. Odejmowanie bliskich sobie wartości prowadzi do znacznego obniżania dokładności obliczeń.

Obecne metody różniczkowania numerycznego wykorzystują aproksymację lub interpolację. Gdy funkcja nie zawiera zaburzeń w postaci szumu, interpoluje się ją najczęściej za pomocą wielomianów. Przy odpowiednio małych odległościach między węzłami wyniki są zadowalające. Zaś gdy różniczkowany sygnał zawiera szum, wykorzystywana jest aproksymacja (najczęściej wielomianowa punktowa średniokwadratowa) do lokalnego opisu różniczkowanej funkcji. Funkcję tę później można różniczkować. Zatem rozważane różniczkowanie numeryczne sprowadza się do różniczkowania wielomianu aproksymującego lub interpolującego wybrany podzbiór punktów funkcji różniczkowanej. Innymi słowy, znając zbiór węzłów x_i i wartości funkcji y_i poszukuje się wartości

$$y'_i = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_i} \tag{5.1}$$

pochodnej funkcji w zadanych wcześniej węzłach.

Większość metod różniczkowania polega na interpolacji dyskretnej funkcji różniczkowanej inną funkcją, która najczęściej jest wielomianem. Następnie dokonuje się różniczkowania uzyskanej funkcji. Jak wskazano w rozdziale na temat interpolacji, powszechność stosowania wielomianów wynika z twierdzenia Weierstrassa, mówiącego o tym, że można z dowolną dokładnością aproksymować dowolną funkcję za pomocą wielomianów.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Poniżej zostaną przedstawione przykładowe sposoby przybliżonego wyznaczania wartości pochodnej funkcji danej w postaci dyskretnej.

5.1. Wyznaczanie wartości pochodnej

Podstawowe zależności pozwalające na przybliżone wyznaczenie pochodnej można otrzymać wprost z definicji pochodnej funkcji w punkcie, pomijając przejście graniczne, a przyjmując założenie o niewielkim h, co daje

$$f'(x_i) = \lim_{h\to 0} \frac{f(x_i+h) - f(x_i)}{h} \approx \frac{f(x_i+h) - f(x_i)}{h} \tag{5.2}$$

Pomijając inne wady tej metody, można stwierdzić, że jedną z podstawowych jest fakt, iż długość tak otrzymanego przybliżenia pochodnej jest mniejsza o jedną wartość niż danej wejściowej, co zazwyczaj nie jest wygodne. Poza tym, aby oszacowanie wielkości pochodnej było dokładne, użyty krok musi mieć małą długość, co prowadzi do niewielkich przyrostów wartości funkcji, zaś odejmowanie bliskich sobie wartości, jak to zostało wcześniej wspomniane, powoduje katastrofalny wzrost względnych błędów zaokrągleń. Jeszcze inną słabością postępowania przedstawionego w zależności (5.2) jest brak jakiejkolwiek informacji o wielkości popełnianego błędu.

Zależność (5.2) można też otrzymać z rozwinięcia funkcji $f(x)$ w szereg Taylora w przód z krokiem h wokół punktu x_i

$$f(x_i+h) = f(x_i) + hf'(x_i) + \frac{h^2}{2!}f''(x_i) + \frac{h^3}{3!}f'''(x_i) + \cdots \tag{5.3}$$

następnie ograniczając się do wyrazów rozwinięcia do pierwszej pochodnej

$$f(x_i+h) \approx f(x_i) + hf'(x_i) + \mathcal{O}(h^2) \tag{5.4}$$

oraz wyznaczając pierwszą pochodną, otrzymuje się oszacowanie błędu tej aproksymacji

$$hf'(x_i) = f(x_i+h) - f(x_i) + \mathcal{O}(h^2) \tag{5.5}$$

czyli

$$f'(x_i) = \frac{f(x_i+h) - f(x_i)}{h} + \mathcal{O}(h) \tag{5.6}$$

W zależnościach (5.4), (5.5) oraz (5.6) pojawiło się wyrażenie $\mathcal{O}(h^2)$, które oznacza tyle, że reszta (obcięcia) szeregu Taylora jest rzędu h^2. Ten zapis należy rozmieć tak, że popełniany błąd jest proporcjonalny do h^2. Ponieważ h jest zazwyczaj małe, wartości kolejnych potęg kroku h stają się coraz mniejsze, jednocześnie mianownik bardzo szybko rośnie, wobec tego mają coraz mniejszy wpływ na wielkość błędu wynikającego z obcięcia szeregu Taylora. Zatem decydujący wpływ na wielkość błędu ma pierwszy pomijany składnik szeregu.

Ponieważ zachodzi dzielenie przez h, pogarsza się dokładność metody. Zamiast rzędu drugiego uzyskuje się metodę rzędu pierwszego, co skutkuje błędami proporcjonalnymi do długości kroku. Zależność (5.6) przedstawia różniczkowanie w przód.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Oprócz wskazanego sposobu, różniczkowanie można realizować w tył lub też centralnie, co zostanie wykazane poniżej. Jeżeli rozwinięcie w szereg Taylora odbędzie się w tył

$$f(x_i - h) = f(x_i) - hf'(x_i) + \frac{h^2}{2!}f''(x_i) - \frac{h^3}{3!}f'''(x_i) + \cdots \tag{5.7}$$

to ograniczając się do rozwinięcia z pierwszą pochodną włącznie, można napisać

$$f(x_i - h) = f(x_i) - hf'(x_i) + \mathcal{O}(h^2)$$

co prowadzi do oszacowania

$$hf'(x_i) = f(x_i) - f(x_i - h) + \mathcal{O}(h^2) \tag{5.8}$$

a po podzieleniu przez krok h otrzymuje się następujące oszacowanie wartości pochodnej w punkcie

$$f'(x_i) = \frac{f(x_i) - f(x_i - h)}{h} + \mathcal{O}(h) \tag{5.9}$$

Inną zależność można uzyskać po odjęciu stronami równań (5.3) oraz (5.7) co, po obcięciu za pierwszą pochodną, prowadzi do zależności, której błąd jest rzędu h^3, ponieważ drugie pochodne znikają

$$f(x_i + h) - f(x_i - h) = 2hf'(x_i) + \mathcal{O}(h^3)$$

Dzięki tej zależności można zapisać pierwszą pochodną jako

$$f'(x_i) = \frac{f(x_i + h) - f(x_i - h)}{2h} + \mathcal{O}(h^2) \tag{5.10}$$

Otrzymane w ten sposób wzory mają różne właściwości numeryczne. Zależności (5.6) oraz (5.9) dają błąd proporcjonalny do długości kroku, ponieważ prosta aproksymująca pochodną jest prowadzona z lewego lub prawego punktu analizowanego przedziału. Zależność (5.10) daje błąd proporcjonalny do kwadratu kroku. Ten efekt jest skutkiem prowadzenia prostej aproksymującej przez punkty skrajne przedziału otaczającego punkt, w którym jest wyznaczana pochodna. Opisane metody zostały pokazane na rysunku 5.1. Widać, iż tylko iloraz różnicowy centralny daje pochodną, której wartość jest zbliżona do wartości dokładnej. Ilorazy różnicowe w przód i w tył wymagają małego kroku, by mogły być uznane za sensowne estymatory pochodnej funkcji w punkcie.

Podobnie można wyznaczyć przybliżoną wartość drugiej pochodnej. Dodając stronami równania (5.3) oraz (5.7), otrzymuje się zależność

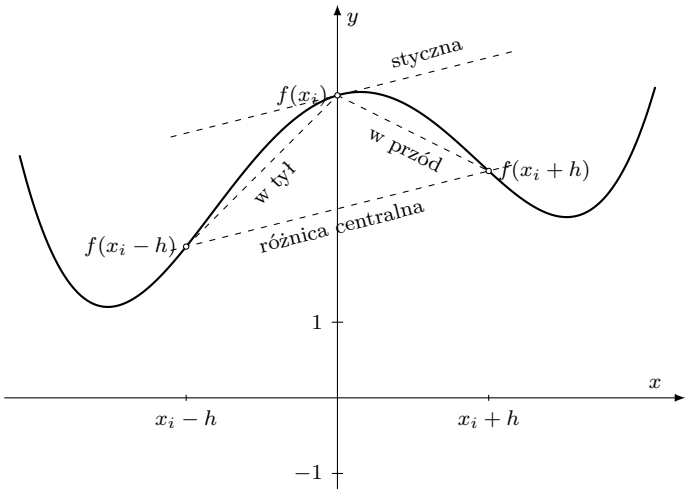
$$f(x_i + h) + f(x_i - h) = 2f(x_i) + 2\frac{h^2}{2!}f''(x_i) + \mathcal{O}(h^4)$$

co pozwala wyznaczyć drugą pochodną jako

$$f''(x_i) = \frac{f(x_i + h) - 2f(x_i) + f(x_i - h)}{h^2} + \mathcal{O}(h^2) \tag{5.11}$$

Jak pokazano powyżej, rozwinięcie funkcji w szereg Taylora pozwala budować zależności umożliwiające wyznaczenie pochodnej w sposób przybliżony. Należy przy tym pamiętać, iż rozwinięcie funkcji w szereg Taylora oznacza aproksymację wielomianem odpowiedniego stopnia.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 5.1. Ilorazy różnicowe jako numeryczne przybliżenia pochodnej

5.1.1. Różniczkowanie trójpunktowe

Jako rozwinięcie przedstawionego w poprzednim podrozdziale postępowania, zostanie przedstawiony sposób uzyskiwania jeszcze lepszych zależności określających pochodną funkcji danej w postaci dyskretnych punktów.

Rozwijając funkcję w szereg Taylora przy koku h oraz $2h$, otrzymuje się

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f'''(x_0) + \cdots \tag{5.12}$$

$$f(x_0+2h) = f(x_0) + 2hf'(x_0) + \frac{(2h)^2}{2!}f''(x_0) + \frac{(2h)^3}{3!}f'''(x_0) + \cdots \tag{5.13}$$

Mnożąc przez 4 równanie (5.12) i odejmując od niego równanie (5.13) a następnie obcinając wynik za pierwszą pochodną, można uzyskać

$$4f(x_0+h) - f(x_0+2h) = 3f(x_0) + 2hf'(x_0) + \mathcal{O}(h^3).$$

Zatem pierwsza pochodna wyraża się zależnością

$$f'(x_0) = \frac{1}{2h}\left[-3f(x_0) + 4f(x_0+h) - f(x_0+2h)\right] + \mathcal{O}(h^2) \tag{5.14}$$

Otrzymany wzór aproksymuje pierwszą pochodną funkcji w punkcie, z błędem proporcjonalnym do kwadratu kroku.

Podobną zależność można uzyskać, rozwijając w szereg Taylora funkcję w tył, czyli $f(x-h)$ oraz $f(x-2h)$, jak zostało to przedstawione poniżej.

$$f(x_i-h) = f(x_i) - hf'(x_i) + \frac{h^2}{2!}f''(x_i) - \frac{h^3}{3!}f'''(x_i) + \cdots \tag{5.15}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$f(x_i - 2h) = f(x_i) - 2hf'(x_i) + \frac{(2h)^2}{2!}f''(x_i) - \frac{(2h)^3}{3!}f'''(x_i) + \cdots \tag{5.16}$$

Po obcięciu do wyrazów zawierających drugą pochodną włącznie, należy pomnożyć równanie (5.15) przez 4 i odjąć równanie (5.16), co z kolei da następujący wynik

$$4f(x_i - h) - f(x_i - 2h) = 3f(x_i) - 2hf'(x_i) + \mathcal{O}(h^3).$$

Zatem pierwsza pochodna w zależności od punktów poprzednich, może być zapisana jako

$$f'(x_i) = \frac{1}{2h}\left[3f(x_i) - 4f(x_i - h) + f(x_i - 2h)\right] + \mathcal{O}(h^2) \tag{5.17}$$

Podobnie jak poprzednio, otrzymany wzór aproksymuje pierwszą pochodną w punkcie, z błędem proporcjonalnym do kwadratu kroku.

Zależności opisujące metodę trójpunktową różniczkowania numerycznego można również wyprowadzić, dokonując interpolacji wielomianowej funkcji składającej się z trzech punktów, a następnie wyznaczyć pochodną otrzymanego wielomianu. Aby lepiej zrozumieć przedstawione tu wyprowadzenie, niezbędne może być sięgnięcie do rozdziału 2, omawiającego zagadnienie interpolacji Lagrange'a.

Wyprowadzenie za pomocą interpolacji. Niech będą dane trzy równoodległe punkty (x_0, y_0), (x_1, y_1), (x_2, y_2) (patrz rysunek 5.2). Gdy wykorzysta się interpolację Lagrange'a, wielomiany składowe dla poszczególnych punktów będą wyrażały się zależnościami

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2}$$

$$l_1(x) = \frac{x - x_0}{x_1 - x_0}\frac{x - x_2}{x_1 - x_2}$$

$$l_2(x) = \frac{x - x_0}{x_2 - x_0}\frac{x - x_1}{x_2 - x_1}$$

zaś wynikowy wielomian interpolacyjny może być zapisany wzorem

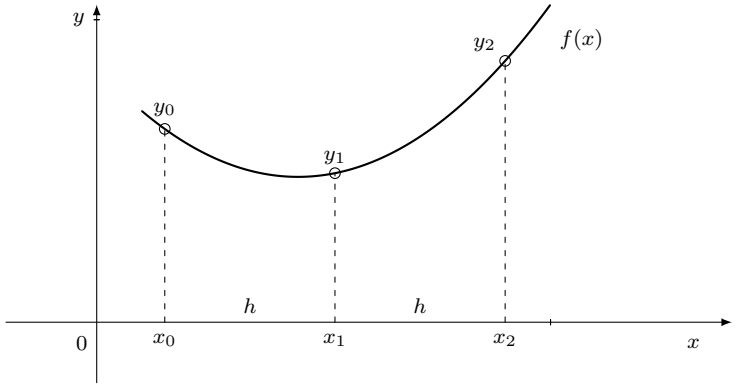
$$p(x) = y_0 l_0 + y_1 l_1 + y_2 l_2 = y_0\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + \\ + y_2\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \tag{5.18}$$

natomiast jego pochodna względem x jest określona w sposób następujący

$$p'(x) = \frac{x_0^2 y_1 - x_1^2 y_0 - x_0^2 y_2 + x_2^2 y_0 + x_1^2 y_2 - x_2^2 y_1 - 2xx_0 y_1 + 2xx_1 y_0 + 2xx_0 y_2}{(x_0 - x_1)(x_0 - x_2)(x_1 - x_2)} - \\ + \frac{-2xx_2 y_0 - 2xx_1 y_2 + 2xx_2 y_1}{(x_0 - x_1)(x_0 - x_2)(x_1 - x_2)} \tag{5.19}$$

Dzięki wyznaczeniu tejże pochodnej można określić zależności na wartość pochodnej funkcji interpolującej założony na początku zbiór trzech punktów.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 5.2. Metoda trójpunktowa – wyprowadzenie za pomocą interpolacji

Wyznaczając wartość pochodnej w punkcie $x = x_0$ i uwzględniając równe odległości h między węzłami, można uzyskać zależność

$$p'(x_0) = \underbrace{\frac{y_0}{x_0 - x_1}}_{-h} + \underbrace{\frac{y_0}{x_0 - x_2}}_{-2h} - \underbrace{\frac{y_1 \overbrace{(x_0 - x_2)}^{-2h}}{(x_0 - x_1)(x_1 - x_2)}}_{-h \quad -h} + \underbrace{\frac{y_2 \overbrace{(x_0 - x_1)}^{-h}}{(x_0 - x_2)(x_1 - x_2)}}_{-2h \quad -h} =$$

$$= \frac{-y_0}{h} - \frac{y_0}{2h} + \frac{2y_1}{h} - \frac{y_2}{2h} = \frac{1}{2h}\left[-3y_0 + 4y_1 - y_2\right] \quad (5.20)$$

Podobnie wartość pochodnej wielomianu w punkcie środkowym $x = x_1$, po uwzględnieniu jednakowych odległości h między punktami, przedstawia się następująco

$$p'(x_1) = \underbrace{\frac{y_1}{x_1 - x_2}}_{-h} - \underbrace{\frac{y_1}{x_0 - x_1}}_{-h} + \underbrace{\frac{y_0 \overbrace{(x_1 - x_2)}^{-h}}{(x_0 - x_1)(x_0 - x_2)}}_{-h \quad -2h} - \underbrace{\frac{y_2 \overbrace{(x_0 - x_1)}^{-h}}{(x_0 - x_2)(x_1 - x_2)}}_{-2h \quad -h} =$$

$$= \frac{1}{2h}\left[y_2 - y_0\right] \quad (5.21)$$

Dla ostatniego punktu $x = x_2$ wartość pochodnej po uproszczeniach, będzie się przedstawiać następująco

$$p'(x_2) = \underbrace{\frac{y_1 \overbrace{(x_0 - x_2)}^{-2h}}{(x_0 - x_1)(x_1 - x_2)}}_{-h \quad -h} - \underbrace{\frac{y_2}{x_1 - x_2}}_{-h} - \underbrace{\frac{y_0 \overbrace{(x_1 - x_2)}^{-h}}{(x_0 - x_1)(x_0 - x_2)}}_{-h \quad -2h} - \underbrace{\frac{y_2}{x_0 - x_2}}_{-2h} =$$

$$= \frac{1}{2h}\left[y_0 - 4y_1 + 3y_2\right] \quad (5.22)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Dla ujednolicenia otrzymanych wzorów z poprzednio wyprowadzonymi, zostanie dokonana zmiana formy zapisu

$$y_i = f(x_i) \quad \text{dla } i = 0, 1, 2, \ldots, n$$

oraz zostanie przyjęte, że punkt x_1 odpowiada dowolnemu punktowi wewnątrz przedziału różniczkowania $[x_0, x_n]$, zaś ostatniej wartości będzie odpowiadać y_2

$$f(x_n) = f(x_2)$$

Łatwo można się przekonać, iż po tych przekształceniach wynikowe wzory są identyczne z uzyskanymi wcześniej dzięki rozwijaniu funkcji w szereg Taylora z krokiem w przód i tył.

Pojawia się pytanie, czy któraś z metod wyprowadzania tych wzorów jest lepsza? Odpowiedź zależy od kryteriów oceny. Metoda z interpolacją Lagrange'a jest nieco bardziej pracochłonna, ale z dydaktycznego punktu widzenia jest lepsza, ponieważ pozwala uwypuklić zastosowaną interpolację wielomianem drugiego stopnia. Z kolei szereg Taylora wprost daje oszacowanie błędu aproksymacji pochodnej.

W podobny sposób można konstruować zależności pozwalające różniczkować z dużą dokładnością funkcje dane w postaci dyskretnej. Nie wolno jednak dać się zwieść pozornej łatwości takiego różniczkowania. W przypadku danych zaburzonych zakłóceniami wyniki będą bardzo niedokładne.

Metoda trójpunktowa. Po zebraniu powyżej wyprowadzonych zależności (5.20), (5.21) i (5.22) uzyskuje się trzy zależności zwane metodą trójpunktową (5.23).

$$\begin{aligned} f'(x_0) &= \frac{1}{2h}[-3f(x_0) + 4f(x_1) - f(x_2)] \\ f'(x_i) &= \frac{1}{2h}[f(x_{i+1}) - f(x_{i-1})] \quad i = 1, 2, 3, \ldots, n-1 \\ f'(x_n) &= \frac{1}{2h}[f(x_{n-2}) - 4f(x_{n-1}) + 3f(x_n)] \end{aligned} \tag{5.23}$$

Zazwyczaj obliczenia wykonuje się zgodnie z indeksami przy argumentach po lewej stronie znaku równości, czyli pierwszą wartość pochodnej na podstawie pierwszego równania, ostatnią na podstawie ostatniego równania, zaś pochodną w punktach pośrednich oblicza się z zależności drugiej.

Jako ćwiczenie czytelnik może się podjąć zbudowania zależności na pochodną przy wykorzystaniu wielomianu interpolacyjnego przechodzącego przez pięć lub siedem punktów.

Przykład obliczeniowy

Sformułowanie problemu. Należy porównać metody różniczkowania numerycznego dla funkcji danych w postaci dyskretnego zbioru punktów. Porównanie przeprowadzić dla wzorów z błędem proporcjonalnym do długości kroku h i kwadratu długości kroku – h^2. Określić wielkość błędu względnego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązanie. Tabele 5.1 oraz 5.2 zawierają dane, oraz wyniki obliczeń. W pierwszych dwóch kolumnach znajduje się funkcja w postaci dyskretnej. W obu tabelach zamieszczono tę samą funkcję.

Tabela 5.1. Pochodna wyznaczona metodą trójpunktową

x	y	Numeryczne y'	Analityczne y'	$\lvert\varepsilon\rvert$ [%]
0,00000	0,00000	1,02230	1,00000	2,2
0,26180	0,25882	0,95493	0,96593	1,1
0,52360	0,50000	0,85617	0,86603	1,1
0,78540	0,70711	0,69906	0,70711	1,1
1,04720	0,86603	0,49431	0,50000	1,1
1,30900	0,96593	0,25587	0,25882	1,1
1,57080	1,00000	0,00443	0,00000	–

Tabela 5.2. Pochodna wyznaczona za pomocą zależności (5.6), (5.10) oraz (5.9)

Funkcja $y(x)$		W przód		Centralnie		W tył	
x	y	y'	$\lvert\varepsilon\rvert$ [%]	y'	$\lvert\varepsilon\rvert$ [%]	y'	$\lvert\varepsilon\rvert$ [%]
0,00000	0,00000	0,98862	1,1	–	–	–	–
0,26180	0,25882	0,92124	4,6	0,95493	1,1	0,98862	2,3
0,52360	0,50000	0,79109	8,7	0,85617	1,1	0,92124	6,4
0,78540	0,70711	0,60702	14,2	0,69906	1,1	0,79109	11,9
1,04720	0,86603	0,38159	23,7	0,49431	1,1	0,60702	21,4
1,30900	0,96593	0,13015	49,7	0,25587	1,1	0,38159	47,4
1,57080	1,00000	–	–	–	–	0,13015	–

Uważny czytelnik zapewne zorientuje się, że wartości różniczkowanej funkcji są wartościami funkcji $y = \sin(x)$. Funkcja sinus jest bardzo regularna, ma wszystkie

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

potrzebne pochodne i występuje często w problemach inżynierskich. Stosunkowo duże błędy przy różniczkowaniu tak eleganckiej funkcji wskazują na słabości metod różniczkowania numerycznego.

Skoro przy stosunkowo małej długości kroku błędy przy stosowaniu ilorazów różnicowych w przód i tył, osiągają wartości rzędu 50%, pojawia się pytanie, co się stanie, gdy funkcja różniczkowana będzie znacznie szybciej rosnąć niż funkcja sinus. Za przykład niech posłuży funkcja tangens.

5.1.2. Przykład obliczeniowy

Sformułowanie problemu Wyznaczyć wartości pochodnej funkcji $y = \operatorname{tg} x$ w równoodległych punktach przy kroku $h = 0{,}2$ w przedziale $x \in \langle 0; 1{,}4\rangle$. Określić wielkość błędu względnego. Obliczenia przeprowadzić za pomocą metody trójpunktowej.

Rozwiązanie. Rozwiązanie zamieszczono w tabeli 5.3. Można zauważyć większą wartość błędu na lewym brzegu przedziału różniczkowania. Zazwyczaj wartości na brzegu przedziału, wyznaczane za pomocą pierwszej i ostatniej zależności w równaniach (5.23) są obarczone większym błędem niż wartości wyznaczane za pomocą ilorazu centralnego. W analizowanym przykładzie ta zależność nie jest spełniona na prawym brzegu.

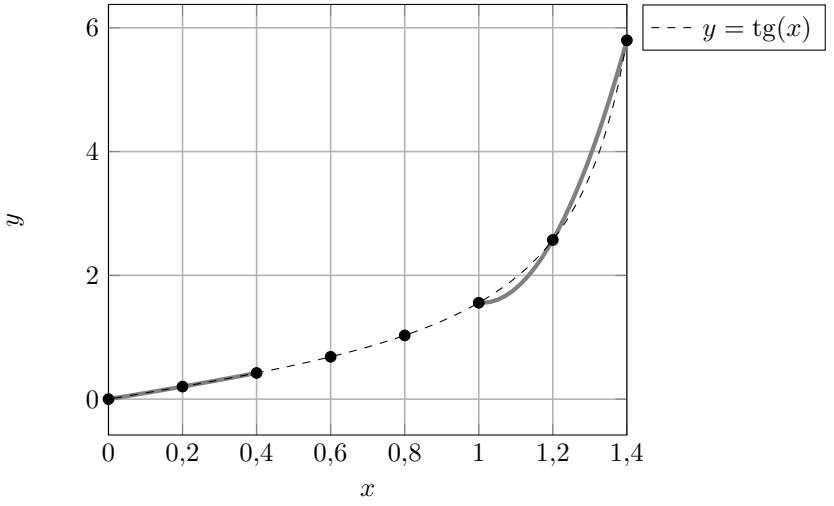
Tabela 5.3. Numeryczne wartości $(\operatorname{tg} x)'$

x	y	y'_{przybl}	y'_{dok}	$\lvert\varepsilon\rvert$ [%]
0,0	0,00000	0,97012	1,00000	3,0
0,2	0,20271	1,05698	1,04109	1,5
0,4	0,42279	1,20357	1,17875	2,1
0,6	0,68414	1,51711	1,46804	3,3
0,8	1,02964	2,18318	2,06016	6,0
1,0	1,55741	3,85628	3,42552	12,6
1,2	2,57215	10,60119	7,61596	39,2
1,4	5,79788	21,65613	34,61546	37,4

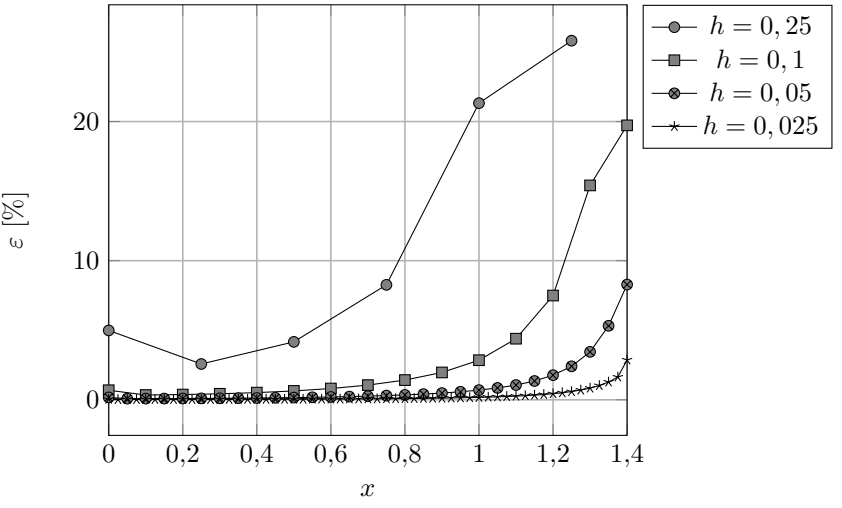
Rysunek 5.3 pozwala zrozumieć dlaczego tak się dzieje. Parabola przebiegająca w przedziale $x \in \langle 1; 1{,}4\rangle$, której pochodne w zaczernionych punktach są aproksymacjami pochodnych funkcji tangens, nie jest dostatecznie dobrą aproksymacją róż-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

niczkowanej funkcji. Zaś z lewej strony przedziału, parabola bardzo dobrze opisuje różniczkowaną funkcję.



Rysunek 5.3. Przebieg różniczkowanej funkcji z parabolami aproksymującymi skrajne fragmenty



Rysunek 5.4. Wielkość błędu względnego przy różniczkowaniu funkcji $y = \operatorname{tg} x$ dla różnych długości kroków

Dla uzupełnienia, rysunek 5.4 ilustruje zmianę wielkości popełnianego błędu względnego przy różniczkowaniu funkcji tangens z coraz mniejszą długością kroku h. Wi-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

doczna jest silna zależność wielkości błędu od długości kroku. Widać też pogorszenie jakości aproksymacji pochodnej w obszarach szybszej zmiany wartości funkcji.

5.2. Inne metody różniczkowania

Podobnie jak dla metody trójpunktowej, można wyprowadzić zależności dla większej liczby punktów. Poniżej zostaną podane zależności pozwalające na wyznaczenie pierwszej pochodnej na podstawie czterech (5.24) i pięciu punktów (5.25) [40].

$$\begin{aligned} f'(x_0) &= \frac{1}{6h}[-11y_0 + 18y_1 - 9y_2 + 2y_3] \\ f'(x_1) &= \frac{1}{6h}[-2y_0 - 3y_1 + 6y_2 - y_3] \\ f'(x_2) &= \frac{1}{6h}[y_0 - 6y_1 + 3y_2 + 2y_3] \\ f'(x_3) &= \frac{1}{6h}[-2y_0 + 9y_1 - 18y_2 + 11y_3] \end{aligned} \tag{5.24}$$

$$\begin{aligned} f'(x_0) &= \frac{1}{12h}[-25y_0 + 48y_1 - 36y_2 + 16y_3 - 3y_4] \\ f'(x_1) &= \frac{1}{12h}[-3y_0 - 10y_1 + 18y_2 - 6y_3 + y_4] \\ f'(x_2) &= \frac{1}{12h}[y_0 - 8y_1 + 8y_3 - y_4] \\ f'(x_3) &= \frac{1}{12h}[-y_0 + 6y_1 - 18y_2 + 10y_3 + 3y_4] \\ f'(x_4) &= \frac{1}{12h}[3y_0 - 16y_1 + 36y_2 - 48y_3 + 25y_4] \end{aligned} \tag{5.25}$$

Oczywiście możliwe jest wyprowadzanie wzorów na drugą pochodną funkcji w postaci dyskretnej na przykład za pomocą interpolacji. Podobnie można wykorzystać jeszcze większą liczbę punktów do interpolacji funkcji różniczkowanej. Jak widać z powyższych wzorów, w miarę wzrostu liczby punktów, za pomocą których wyznacza się różniczkowany wielomian, wzory stają się coraz bardziej skomplikowane.

5.2.1. Różniczkowanie sygnałów zaszumionych

Często zdarza się konieczność wyznaczenia pochodnej sygnału cyfrowego. Niestety najczęściej jest to zaszumiony sygnał. Czyli wartości w poszczególnych chwilach czasowych nie są dokładnymi wartościami funkcji różniczkowanej, ale zawierają składniki o charakterze losowym. Wobec tego stosowanie przedstawionych w tym rozdziale wzorów nie jest praktyczne, ponieważ prowadzi do dużych błędów.

W związku z tym zamiast stosować interpolację do wyznaczania wielomianu poddawanego później różniczkowaniu, stosuje się aproksymację, a otrzymany wielomian się różniczkuje i w ten sposób wyznacza się pochodną. Cała rodzina filtrów (procedur numerycznych) realizujących te operacje jest nazywana filtrami Sawitzky'ego–Golaya.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

5.3. Podsumowanie

Różniczkowanie numeryczne jest doskonałą ilustracją zjawiska utraty precyzji przy odejmowaniu. Ponieważ występuje odejmowanie bliskich sobie wartości, więc i utrata precyzji jest znacząca. Im mniejszy krok różniczkowania, tym bliższe stają się odejmowane wartości i tym większa utrata precyzji.

Przedstawione przykłady dotyczyły wybranych funkcji – bardzo regularnych, różniczkowalnych i bez uwzględniania zakłóceń, a mimo to różniczkowanie numeryczne nie było szczególnie dokładną operacją. Uzyskane wielkości błędów świadczą o złym uwarunkowaniu problemu różniczkowania. Utrata informacji o przebiegu funkcji, zachodząca na skutek dyskretyzacji, powoduje dużą niepewność wyznaczenia pochodnej. To zjawisko jest szczególnie istotne przy różniczkowaniu danych empirycznych, gdzie niewielkie zakłócenia, niewpływające na wynik całkowania, mogą uniemożliwić wyznaczenie wartości pochodnej. Z tego powodu rozwinięto wiele innych metod służących do różniczkowania sygnałów cyfrowych, jak na przykład metoda Lanczosa.

Najczęściej zależności wyprowadzone przy różniczkowaniu numerycznym są wykorzystywane przy metodzie różnic skończonych, służącej do rozwiązywania równań różniczkowych. Dzięki tym zależnościom można przeprowadzić dyskretyzację równań różniczkowych, co prowadzi do układu równań algebraicznych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 6

Rozwiązywanie równań różniczkowych

Uznaje się, że rachunek różniczkowy w znanej nam formie powstał w XVII wieku. To właśnie wtedy Izaak Newton oraz Gottfried Wilhelm Leibniz niezależnie rozwinęli rachunek różniczkowy (oczywiście całkowy również) i przedstawili go w postaci bliskiej dzisiejszej. Chociaż ich prace w części stanowiły kontynuacje wysiłków Galileo Galileiego zwanego Galileuszem i Johannesa Keplera, to właśnie owych wielkich badaczy uznaje się za twórców rachunku różniczkowego. Podobnie jak przy wielu innych zagadnieniach matematycznych, motywacji do zajęcia się tymi zagadnieniami dostarczało badanie zjawisk fizycznych, będące nieustającym źródłem coraz to nowych i często bardzo złożonych problemów.

6.1. Podstawowe informacje o równaniach różniczkowych

Równania różniczkowe wykorzystuje się do opisu matematycznego układów fizycznych. Proces tworzenia równań opisujących poszczególne układy jest bardzo szeroko opisany w literaturze. Po podjęciu decyzji dotyczącej wielkości wejściowych i wyjściowych, zazwyczaj zaczyna się od zbudowania tak zwanego modelu fizycznego. Nie wnikając w szczegóły, można stwierdzić, że opracowanie takiego modelu obejmuje następujące kroki:

1. izolacja i dyskretyzacja układu rzeczywistego,
2. pomijanie mało istotnych wpływów,
3. przyjęcie uproszczeń dotyczących układu i jego opisu.

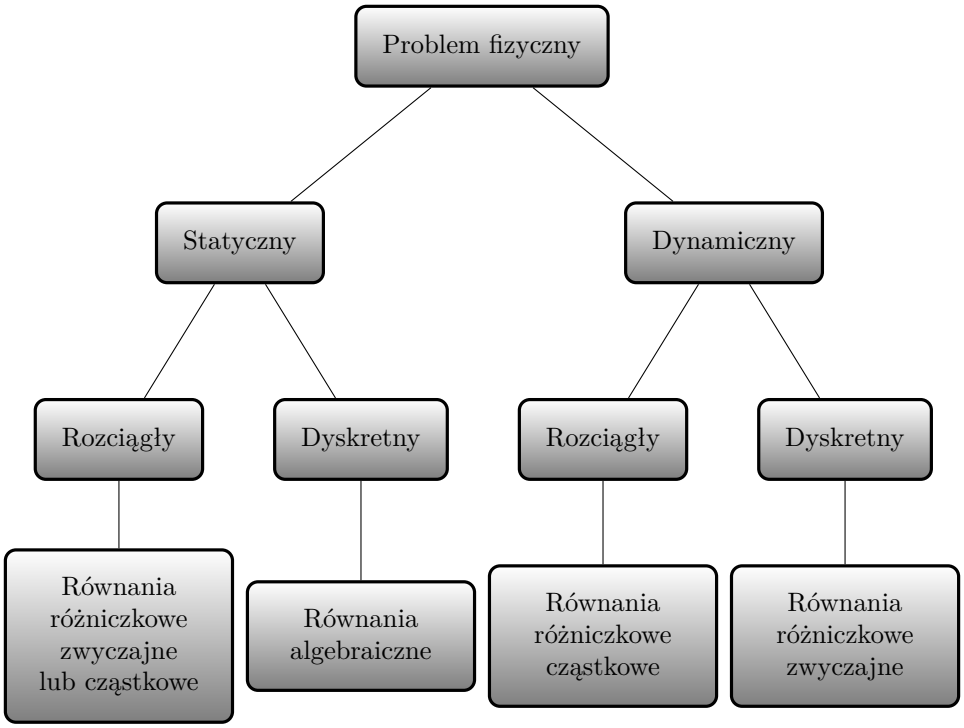
W wyniku zastosowania powyższych uproszczeń uzyska się model fizyczny. Na etapie inżynierskim najczęściej jest to model liniowy, w jasny sposób wiążący wielkości wejściowe i wyjściowe modelu.

Taki model można opisać za pomocą równań. Matematyczny opis modelu fizycznego nazywa się modelem matematycznym. Zatem analiza zachowania układów fizycznych prowadzi do ich opisu matematycznego. W zależności od charakteru badanego zjawiska czy przyjętego modelu fizycznego, do jego opisu stosuje się różne narzędzia matematyczne. Na rysunku 6.1 w uproszczony sposób przedstawiono wpływ charakteru zjawiska na używany opis matematyczny.

Problemy statyczne są opisywane przez równania algebraiczne lub też równania różniczkowe, w których nie występuje zależność od czasu. Uzyskiwane równania różniczkowe mogą być równaniami zwyczajnymi dla układów jednowymiarowych lub

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

cząstkowymi dla układów dwuwymiarowych i trójwymiarowych. Dlatego linia ugięcia belki jest opisywana równaniem różniczkowym zwyczajnym, natomiast ugięcia płyt są opisywane równaniami różniczkowymi cząstkowymi. Warto pamiętać, że na przykład rozkład drgań, kiedy położenie punktów analizowanego obiektu jest zmienne w czasie, też opisuje się równaniami różniczkowymi, w których nie występuje czas. Wynika to z przyjęcia założenia, że przemieszczenia w czasie zmieniają się sinusoidalnie, wobec tego dla danej częstotliwości można badać tylko rozkład amplitud i przesunięć fazowych tych drgań.



Rysunek 6.1. Modele matematyczne stosowane przy opisie problemów fizycznych

Problemy dynamiczne są opisywane równaniami różniczkowymi, w których jedna ze zmiennych niezależnych jest utożsamiana z czasem. W zależności od przyjętego sposobu modelowania układu fizycznego, mogą to być równania różniczkowe cząstkowe lub zwyczajne. Przy dyskretnym modelu układu fizycznego powstają równania różniczkowe zwyczajne. Układy o parametrach rozłożonych opisuje się równaniami różniczkowymi cząstkowymi.

Przy modelowaniu obiektów fizycznych warto zdawać sobie sprawę ze związku między rzędem równania a liczbą magazynów energii w nim występujących. Oznacza to, że rząd równania zależy od liczby niezależnych magazynów energii kinetycznej i potencjalnej. Każdy niezależny magazyn energii dodaje jeden do rzędu równania,

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

dlatego układ masy zawieszonej sprężyście (patrz rysunek 6.6), mający jeden magazyn energii potencjalnej w postaci sprężystości i jeden magazyn energii kinetycznej w postaci masy bezwładnej jest opisywany równaniem różniczkowym zwyczajnym drugiego rzędu.

Klasyfikacja równań różniczkowych. Klasyfikację równań różniczkowych można prowadzić ze względu na różne kryteria. Z punktu widzenia inżyniera, niektóre z tych kryteriów są bardziej istotne, a inne mniej. Poniżej zostaną przedstawione niektóre, uznane przez autorów za najistotniejsze. W zależności od rzędu najwyższej pochodnej mówi się na przykład o równaniach pierwszego rzędu co oznacza, że w równaniu najwyższa jest pierwsza pochodna. Jako przykład można podać równanie rozpadu promieniotwórczego (6.1), podobne do niego równanie opisujące rozładowanie kondensatora czy równanie opisujące konwekcyjną wymianę ciepła z otoczeniem.

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N \tag{6.1}$$

W zależności od problemu, N może oznaczać liczbę cząsteczek materiału promieniotwórczego, napięcie na okładkach kondensatora czy różnicę między temperaturą obiektu i otoczenia. Spośród równań wyższych rzędów, w zagadnieniach technicznych najczęściej spotyka się równania drugiego rzędu, na przykład dobrze znane równanie

$$m\ddot{x} + b\dot{x} + kx = P(t)$$

opisujące ruch układu dyskretnego składającego się z masy m zawieszonej sprężyście na sprężystości k z tłumieniem b. Identyczne równanie opisuje na przykład zachowanie obwodu elektrycznego z elementami odpowiednio L, C i R. Spotyka się też równania wyższych rzędów, chociaż nie są one tak powszechne jak równania pierwszego i drugiego rzędu.

Z kolei w zależności od liczby zmiennych niezależnych można wyróżnić równania różniczkowe zwyczajne, gdy przy opisie problemu wykorzystywana jest jedna zmienna niezależna. Gdy problem opisuje się większą liczbą zmiennych niezależnych, uzyskuje się równania różniczkowe cząstkowe. Pochodne w równaniach różniczkowych zwyczajnych zapisuje się między innymi symbolem $\frac{\mathrm{d}}{\mathrm{d}x}$, y' czy $\dot{y}$, zaś pochodne cząstkowe są zapisywane między innymi jako $\frac{\partial}{\partial x}$. W literaturze można znaleźć kilka innych sposobów zapisu pochodnych, naszym zdaniem mniej czytelnych. Równania różniczkowe mogą być liniowe, quasi-liniowe lub nieliniowe, w zależności od postaci powiązań między poszczególnymi pochodnymi.

Równania różniczkowe cząstkowe rzędu drugiego są dodatkowo dzielone na:

— eliptyczne; można wymienić tutaj równanie Laplace'a

$$\nabla^2 T = 0,$$

opisujące na przykład rozkład temperatur w stanie ustalonym, które po rozwinięciu operatora ∇ przyjmuje postać

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— paraboliczne, dla których typowym przedstawicielem jest równanie dyfuzji (np. ciepła)

$$\frac{\partial T}{\partial t} = \nabla^2 T$$

— hiperboliczne, którego reprezentantem jest równanie falowe

$$\frac{\partial^2 T}{\partial t^2} - c^2 \nabla^2 T = 0$$

opisujące propagację niewielkiego zaburzenia w liniowym ośrodku sprężystym.

Powyższy podział nie jest jedynie podziałem formalnym. Poszczególne klasy równań opisują różne rodzaje zagadnień spotykanych w praktyce inżynierskiej. Zagadnienia techniczne opisywane równaniami różniczkowymi cząstkowymi można zaliczyć do problemów początkowych (przy równaniach różniczkowych zwyczajnych, w których występuje czas jako jedyna zmienna niezależna), problemów brzegowych – najczęściej są to równania cząstkowe przy zagadnieniach ustalonych – oraz problemów początkowo-brzegowych – gdy rozwiązuje się zagadnienia dynamiczne. Oddzielną grupę problemów, które nie będą szczegółowo omawiane w niniejszym skrypcie, stanowią zagadnienia własne. Równania eliptyczne występują w zagadnieniach brzegowych, opisują stany ustalone badanych procesów. Równania paraboliczne i hiperboliczne występują w zagadnieniach początkowo-brzegowych i opisują procesy dynamiczne.

Kilka pojęć podstawowych. Stosując ściśle matematyczne podejście do zagadnienia, przed rozpoczęciem rozwiązywania równania różniczkowego należy stwierdzić, czy równanie ma rozwiązanie oraz czy jest ono jednoznaczne. Zostały udowodnione odpowiednie twierdzenia, podające warunki, jakie musi spełniać równanie różniczkowe, by miało rozwiązanie [23]. Ponieważ w praktyce inżynierskiej zazwyczaj rozwiązuje się problemy realizowalne technicznie, oznacza to, że równania opisujące dane zagadnienie muszą mieć rozwiązanie, jak to wynika z doświadczenia. Zatem problem istnienia rozwiązania nie będzie omawiany. Osoby zainteresowane mogą zgłębić temat, studiując podręczniki poświęcone równaniom różniczkowym, np. [23, 29].

Bardzo ważną właściwością równań różniczkowych jest ciągła zależność rozwiązania od warunków początkowych. Dzięki niej można liczyć na to, że przy dostatecznie małym kroku całkowania uzyskane rozwiązanie jest bliskie rozwiązaniu dokładnemu.

Stabilność jest pojęciem, które można definiować na różne sposoby. W sensie Lapumowa stabilność będzie oznaczać, że dla niewielkich zmian warunków początkowych uzyska się niewielkie zmiany rozwiązania w całym przedziale nieskończonym [29]. Można powiedzieć, że to rozszerzenie ciągłej zależności rozwiązania od warunków początkowych. Jednak chociaż znane warunki istnienia i jednoznaczności rozwiązania równania różniczkowego zapewniają ciągłość rozwiązania w zależności od warunków początkowych, to jednak nie gwarantują stabilności. Niemniej – odwołując się do praktyki – ponieważ rozwiązywane przez inżynierów równania różniczkowe opisują najczęściej procesy stabilne, można oczekiwać, iż równania różniczkowe stanowiące adekwatny opis takich procesów też będą stabilne.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Do badania stabilności metod rozwiązywania równań różniczkowych wykorzystuje się równanie testowe w postaci

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \lambda y, \quad y(0) = 1 \tag{6.2}$$

gdzie λ jest zespoloną stałą z ujemną częścią rzeczywistą. Dla $x \to \infty$ rozwiązanie $y(x) \to 0$. Jeżeli rozwiązanie tego równania dla stałego dodatniego kroku h osiąga wartość 0, gdy $x \to \infty$, to mówi się, że metoda jest A-stabilna [11].

Zbieżność metody numerycznej rozwiązywania równań różniczkowych można ująć następująco: wraz ze zmniejszeniem długości kroku całkowania h rozwiązywanie przybliżone powinno być zbieżne do rozwiązania dokładnego.

Błędy występujące podczas rozwiązywania równań różniczkowych zasadniczo można podzielić na błędy lokalne i globalne. Zazwyczaj błąd globalny obcięcia (albo inaczej błąd metody) jest o rząd wielkości większy niż błąd lokalny. Wynika to z konieczności wykonania wielu kroków całkowania, by uzyskać przybliżone rozwiązanie równania różniczkowego w zadanym przedziale. Zatem metoda dająca błąd lokalny rzędu $\mathcal{O}(h^3)$ daje błąd globalny rzędu $\mathcal{O}(h^2)$.

Podczas rozwiązywania równań różniczkowych wielkość popełnianych błędów zależy od długości kroku całkowania h. Przy analizie metody numerycznej służącej do rozwiązywania równań różniczkowych rozwija się ten błąd $r_n(h)$ (błąd metody) w szereg Taylora, w taki sposób jak w równaniu (6.3).

$$r_n(h) = r_n(0) + hr'_n(0) + \frac{h^2}{2!}r''_n(0) + \frac{h^3}{3!}r'''_n(0) + \ldots \tag{6.3}$$

Zazwyczaj wymaga się, by błąd był najmniejszy. Dlatego dąży się do tego, by w błędzie metody występowały tylko wyższe pochodne. Z tym wiąże się pojęcie rzędu metody [29] rozwiązywania równań różniczkowych. Mówi się, że metoda jest rzędu p, jeżeli w zależności (6.3) na $r_n(h)$ nie występują pochodne rzędu do p włącznie. W związku z tym błąd popełniany w każdym kroku (błąd lokalny) jest rzędu $\mathcal{O}(h^{p+1})$.

Ponieważ błąd lokalny rozwiązania jest wyrażony wielomianem, w którym występują potęgi długości kroku h, można stosować ekstrapolację Richardsona, by znacząco zmniejszać tych wpływ błędów [5]. Można ją też wykorzystać do oceny lokalnego błędu na potrzeby automatycznego doboru długości kroku całkowania.

Zadania sztywne [5] zwane też równaniami sztywnymi [23] czy problemami sztywnymi stawiają bardzo duże wymagania metodom numerycznym stosowanym do ich rozwiązania. Większość metod jawnych nie sprawdza się przy rozwiązywaniu tego typu problemów. Sama nazwa może być myląca, ponieważ może się kojarzyć z układami mechanicznymi. Sztywność równania oznacza szybkie zanikanie składowych rozwiązania z czasem. I właśnie te zanikające z czasem składowe decydują o długości kroku całkowania takiego równania. Z tego typu zjawiskami można mieć do czynienia przy modelowaniu złożonych układów, w których występują znaczne różnice w stałych czasowych poszczególnych podukładów. Część układu o niewielkiej stałej czasowej daje szybko zanikające składowe rozwiązania, ale i tak trzeba je uwzględniać w doborze

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

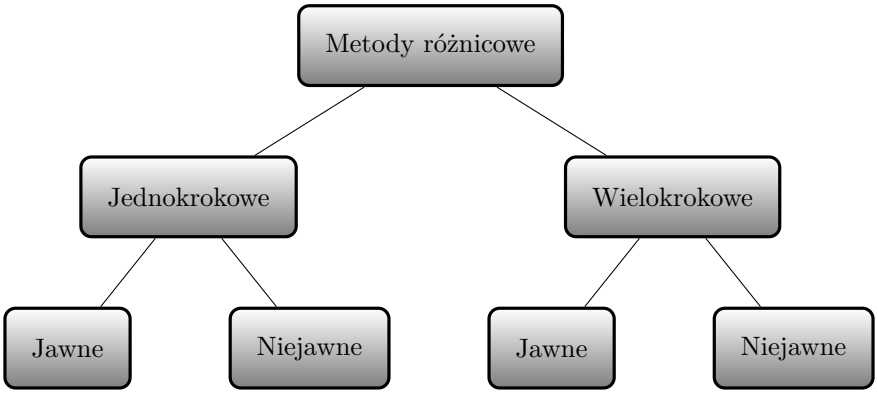
długości kroku całkowania w trakcie całego procesu rozwiązywania. Aby rozwiązywać takie problemy, konieczne staje się używanie niewielkiej długości kroku h, chociaż na pierwszy rzut oka charakter rozwiązania powinien pozwalać na znacznie większą jego długość. Ta sytuacja powoduje, że proces rozwiązywania tego typu równań różniczkowych staje się bardzo nieefektywny, zaś o długości kroku rozwiązania nie decydują wymagania dotyczące dokładności rozwiązania tylko wymaganie stabilności.

6.2. Metody rozwiązywania problemów początkowych

Rozwiązanie równania różniczkowego, jako problem analizy matematycznej, wymaga wyznaczenia funkcji, która spełnia owo równanie wraz z dodatkowymi warunkami początkowymi czy brzegowymi. Takie rozwiązanie może być niezwykle skomplikowane, zwłaszcza gdy obszar rozwiązania i towarzyszące równaniu warunki nie mają prostej postaci. Nawet dla prostych równań różniczkowych proces rozwiązania może być skomplikowany. Z tego powodu w metodach numerycznych wykorzystuje się różne sposoby, by wyjściowe równanie różniczkowe przekształcić tak, by dało się je rozwiązać metodami algebry. Wśród obecnie stosowanych metod, można między innymi wyróżnić dwie następujące grupy:

— metody wykorzystujące szeregi, wśród których można wymienić metodę współczynników nieoznaczonych, metodę kolejnego różniczkowania czy metodę Picarda;
— metody różnicowe (patrz rys. 6.2).

Więcej informacji na temat metod rozwiązywania równań różniczkowych za pomocą szeregów można znaleźć na przykład w publikacji [33].



Rysunek 6.2. Podstawowy podział metod różnicowych rozwiązywania równań różniczkowych

W dalszych rozważaniach zostanie położony nacisk na drugą grupę metod, czyli na metody różnicowe. W metodach tych dokonuje się przekształcenia równania różniczkowego do równania różnicowego, przy czym bardzo istotny staje się krok h owej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

dyskretyzacji. Żąda się by równanie różnicowe było zgodne z równaniem różniczkowym, czyli by przy kroku dążącym do zera, równanie różnicowe dawało rozwiązanie identyczne z rozwiązaniem równania różniczkowego.

Metody różnicowe można podzielić zgodnie ze schematem przedstawionym na rysunku 6.2. W następnej części rozdziału zostaną omówione poszczególne rodzaje metod wraz z przykładami. Ale zanim zostaną omówione poszczególne metody, należy zadać pytanie: co z równaniami różniczkowymi rzędów wyższych niż jeden?

6.2.1. Równania różniczkowe wyższych rzędów

Często poprzez zamianę zmiennych równanie różniczkowe zwyczajne o stałych współczynnikach rzędu n da się przedstawić jako układ n równań różniczkowych rzędu pierwszego. Poniżej w skrócie zostanie przedstawiony tok postępowania przy takiej zamianie zmiennych.

Dane jest równanie różniczkowe zwyczajne rzędu n zapisane w ten sposób, by z lewej strony znajdowała się najwyższa pochodna

$$\frac{\mathrm{d}^n x}{\mathrm{d}t^n} = f\left(t, x, \dot{x}, \ddot{x}, \ldots, \frac{\mathrm{d}^{n-1} x}{\mathrm{d}t^{n-1}}\right) \tag{6.4}$$

Wprowadzając nowe zmienne z_i

$$\begin{aligned} z_1 &= x \\ z_2 &= \dot{z_1} \\ &\vdots \\ z_n &= \dot{z}_{n-1} = \tfrac{\mathrm{d}^{n-1} x}{\mathrm{d}t^{n-1}} \end{aligned} \tag{6.5}$$

można uzyskać układ równań różniczkowych pierwszego rzędu zgodnie z poniższymi zależnościami

$$\begin{cases} \dot{z_1} &= z_2 \\ \dot{z_2} &= z_3 \\ &\vdots \\ \dot{z_n} &= f\left(t, x, \dot{x}, \ddot{x}, \ldots, \tfrac{\mathrm{d}^{n-1} x}{\mathrm{d}t^{n-1}}\right) \end{cases} \tag{6.6}$$

Przedstawiona metoda transformacji równania rzędu n do n równań pierwszego rzędu nie jest jedyną, jaką można zastosować. Nie każde równanie da się w ten sposób przekształcić. Niemniej większość równań spotykanych w praktyce inżynierskiej można w taki sposób sprowadzać do układu równań pierwszego rzędu.

Opisywane dalej metody, służące do rozwiązywani równań pierwszego rzędu, można stosować do rozwiązywania układów równań pierwszego rzędu. Najłatwiej i najwygodniej można to robić, gdy układ równań zostanie zapisany w postaci macierzowej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Przykład

Równanie opisujące ruch układu (patrz rysunek 6.6), składającego się z masy m zawieszonej sprężyście z tłumieniem opisuje równanie

$$m\ddot{x} + b\dot{x} + kx = f(t, x), \tag{6.7}$$

gdzie $f(t, x)$ jest siłą wymuszającą, niezaznaczoną na rysunku, k oznacza współczynnik sprężystość zaś b jest współczynnikiem tłumienia wiskotycznego.

W pierwszym kroku należy po lewej stronie zostawić składnik z najwyższą pochodną i podzielić obie strony równania przez wartość m

$$\ddot{x} = -\frac{b}{m}\dot{x} - \frac{k}{m}x + \frac{f(t, x)}{m}.$$

Zgodnie z równaniem (6.5) dokonuje się podstawień, które w tym przypadku przyjmują postać

$$\begin{aligned} z_1 &= x \\ z_2 &= \dot{z_1} = \dot{x} \end{aligned}$$

Wobec tego, otrzymuje się układ równań pierwszego rzędu, zapisany przy użyciu powyższych podstawień

$$\begin{aligned} \dot{z_1} &= z_2 \\ \dot{z_2} &= -\frac{b}{m}z_2 - \frac{k}{m}z_1 + \frac{f(t, z_1)}{m} \end{aligned}$$

Dla wygody można zapisać go w postaci macierzowej

$$\begin{bmatrix} \dot{z_1} \\ \dot{z_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{f(t,z_1)}{m} \end{bmatrix}$$

Po wprowadzeniu następujących podstawień

$$\frac{\mathrm{d}\mathbb{z}}{\mathrm{d}t} = \begin{bmatrix} \dot{z_1} \\ \dot{z_2} \end{bmatrix}, \quad \mathbb{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix}, \quad \mathbb{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad \mathbb{b} = \begin{bmatrix} 0 \\ \frac{f(t,z_1)}{m} \end{bmatrix} \tag{6.8}$$

równanie przyjmuje postać finalną

$$\frac{\mathrm{d}\mathbb{z}}{\mathrm{d}t} = \mathbb{A}\mathbb{z} + \mathbb{b}$$

Taki układ równań można rozwiązać, stosując dokładnie te same zależności, które służą do rozwiązywania równań różniczkowych pierwszego rzędu. Jedyną niedogodnością z tym związaną jest konieczność stosowania rachunku macierzowego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.2.2. Metody jednokrokowe

Aktualnie w praktyce inżynierskiej metody jednokrokowe są tożsame z metodami Rungego–Kutty. Z tego powodu na nich będzie się skupiał niniejszy podrozdział. W większości przypadków omówione zostaną jawne warianty metod.

Metoda Eulera

Najprostszą metodą rozwiązywania równań różniczkowych zwyczajnych jest metoda Eulera. Co ciekawe, jest ona zaliczana zarówno do grupy metod Rungego–Kutty, jak i do grupy metod wielokrokowych w obu przypadkach, jako ich najprostszy przedstawiciel [29].

Metoda Eulera służy do rozwiązywania równań różniczkowych pierwszego rzędu w postaci

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y), \quad y(x_0) = y_0$$

Warunek początkowy $y(x_0) = y_0$ pozwala na jednoznaczne rozwiązanie takiego równania. Rozwiązanie zostanie zrealizowane w n dyskretnych punktach x_i. Przyjmując oznaczenia

$$y_i = y(x_i), \quad \text{oraz} \quad y_{i+i} = y(x_i + h)$$

można zapisać lewą stronę równania w postaci pochodnej funkcji w punkcie jako

$$\lim_{h \to 0} \frac{y_{i+1} - y_i}{h} = f(x_i, y_i)$$

Z powodu wprowadzonej dyskretyzacji konieczne jest wskazanie po prawej stronie równania punktu, w którym jest wyznaczana wartość funkcji f. Zakładając niewielkie wartości kroku całkowania h, można pominąć przejście do granicy $h \to 0$, co prowadzi do równania różnicowego, odpowiadającego wyjściowemu równaniu różniczkowemu.

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i)$$

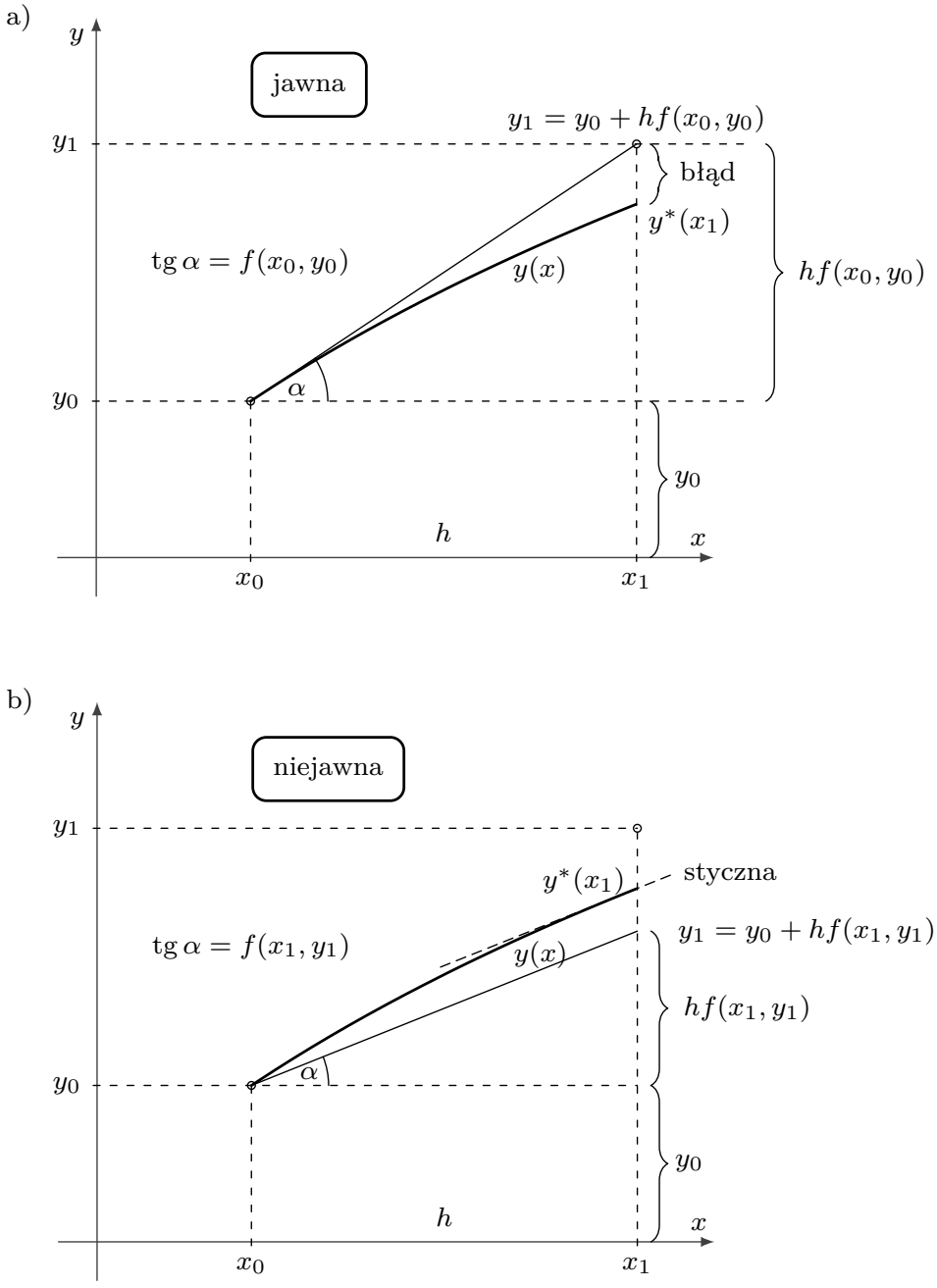
Proste przekształcenie daje jawną postać metody Eulera, która polega na przybliżeniu szukanego rozwiązania odcinkami prostych stycznych do rozwiązania w dyskretnych punktach oddalonych od siebie o długość kroku h (patrz rysunek 6.3)

$$y_{i+1} = y_i + hf(x_i, y_i) \tag{6.9}$$

Metoda Eulera w postaci jawnej wymaga zastosowania niewielkich długości kroku, aby uzyskane wyniki mogły być uznane za dobre przybliżenie rozwiązania równania różniczkowego. Przy większych wartościach h metoda staje się niestabilna. Można, oczywiście, do wyznaczania wartości funkcji $f(x, y)$, bezpośrednio służącej do określania następnego punku rozwiązania, użyć punktu $i + 1$ zamiast i. W takiej sytuacji uzyska się niejawną metodę Eulera. Niejawna postać metody Eulera jest bezwarunkowo stabilna [23], co zostanie pokazane w dalszych rozważaniach, i opisywana jest zależnością

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) \tag{6.10}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 6.3. Interpretacja geometryczna metody Eulera jawnej (a) i niejawnej (b)

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Aby określić wielkości błędu lokalnego w metodzie Eulera, trzeba przyjąć, że punkt (x_0, y_0) należy do dokładnego rozwiązania równania różniczkowego. W takim przypadku, zgodnie z rysunkiem 6.3, błąd można wyznaczyć jako różnicę między wartością y_1 obliczoną metodą Eulera a wartością $y^*(x_1)$, będącą dokładnym rozwiązaniem równania różniczkowego. Wykorzystując rozwinięcie w szereg Taylora, można to rozumowanie zapisać jako

$$y^*(x_1) = y(x_0 + h) = \underbrace{y(x_0) + h \overbrace{\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_0}}^{f(x_0,y_0)}}_{y_1} + \frac{h^2}{2} \left.\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right|_{x=x_0} + \frac{h^3}{3!} \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_0} + \dots \quad (6.11)$$

czyli błąd będący różnicą między rozwiązaniem dokładnym i przybliżonym wynosi

$$y^*(x_1) - y_1 = \frac{h^2}{2} \left.\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right|_{x=x_0} + \frac{h^3}{3!} \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_0} + \dots = \frac{h^2}{2} \left.\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right|_{x=\xi}, \quad \xi \in [x_0, x_1] \quad (6.12)$$

Ponieważ prawa strona równania (6.12) jest resztą szeregu Taylora, jej wartość jest proporcjonalna do h^2. Zatem można powiedzieć, że błąd lokalny metody Eulera jest rzędu $\mathcal{O}(h^2)$, co oznacza, iż w każdym kroku wprowadzany jest błąd proporcjonalny do kwadratu kroku, natomiast globalny błąd jest rzędu $\mathcal{O}(h)$. W związku z tym można powiedzieć, że metoda Eulera jest metodą rzędu pierwszego.

Stabilność

Aby zilustrować sposób analizy stabilności metod rozwiązywania równań różniczkowych, przeanalizowana zostanie stabilność metody Eulera w wersji jawnej oraz niejawnej.

Jawna postać metody Eulera. Jak już wcześniej zostało wspomniane, obszar stabilności określa się dla równania testowego w postaci

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \lambda u, \quad u(t_0) = u_0$$

Dla takiego równania różniczkowego jawną metodę Eulera można zapisać następująco

$$u_{i+1} = u_i + h\lambda u_i$$

co, po prostych przekształceniach, prowadzi do wyrażenia

$$u_{i+1} = u_i + h\lambda u_i = u_i(1 + h\lambda) = u_0(1 + h\lambda)^{i+1}$$

Zatem jawna metoda Eulera dla zadania testowego przyjmuje postać

$$u_{i+1} = u_0 \underbrace{(1 + h\lambda)^{i+1}}_{wsp.\,wzmocnienia}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Przy $i \to \infty$, wyniki pozostają skończone dla współczynnika wzmocnienia spełniającego warunek

$$|1 + h\lambda| \leqslant 1$$

Traktując $h\lambda$ jako zmienną zespoloną z, można narysować nierówność $|1 + z| \leqslant 1$ na płaszczyźnie zespolonej. Rozwiązaniem jest koło o promieniu równym 1 i środku w punkcie $(-1, 0)$. Czyli na osi wartości rzeczywistych obszar rozwiązania tej nierówności jest przedziałem $z \in \langle -2, 0\rangle$. Zatem obszar stabilności jest mocno ograniczony przez długość kroku spełniającą warunek $h \leqslant \frac{2}{\lambda}$.

Niejawna postać metody Eulera. Rozwiązując zadanie testowe jak powyżej za pomocą niejawnej metody Eulera, można napisać następujące równanie

$$u_{i+1} = u_i + h\lambda u_{i+1}$$

co prowadzi do wyrażenia

$$u_{i+1} - h\lambda u_{i+1} = u_i$$

które można przekształcić do postaci

$$u_{i+1} = \frac{u_i}{1 - h\lambda} = \frac{u_0}{(1 - h\lambda)^{i+1}} = u_0 \underbrace{\frac{1}{(1 - h\lambda)^{i+1}}}_{wsp.\,wzmocnienia}$$

Dla $i \to \infty$ wyniki pozostają skończone, gdy współczynnik wzmocnienia spełnia warunek

$$\left|\frac{1}{1 - h\lambda}\right| \leqslant 1$$

Gdy przyjmie się, że $z = h\lambda$, oraz dokona elementarnych przekształceń, powyższą nierówność można zapisać w równoważnej postaci

$$|1 - z| \geqslant 1$$

Obrazem rozwiązania tej nierówności na płaszczyźnie zespolonej jest cała płaszczyzna oprócz wnętrza koła o środku w punkcie $(1, 0)$ i promieniu równym 1. Zatem obszar stabilności obejmuje całą lewą półpłaszczyznę, co oznacza, że dla $Re(\lambda) < 0$ metoda jest bezwarunkowo stabilna.

Przykład obliczeniowy zastosowania jawnej metody Eulera

Sformułowanie problemu. Jawną metodą Eulera rozwiązać równanie różniczkowe pierwszego rzędu z warunkiem początkowym

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -x, \quad x(t_0) = 1$$

Graniczna wielkość kroku całkowania powinna wynosić $h_{max} = 2$. Równanie rozwiązać w przedziale $t \in \langle 0; 6\rangle$. Określić błąd względny rozwiązania.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązanie. Wstawiając do zależności (6.9) prawą stronę rozwiązywanego równania, otrzymuje się następujące rekurencyjne równanie różnicowe z warunkiem początkowym

$$x_{i+1} = x_i + h(-x_i) = x_i - hx_i = x_i(1-h), \quad x(0) = 1$$

Poniższa tabela 6.1 przedstawia uzyskane wartości rozwiązania i wartości błędu dla długości kroku h równych kolejno h_{max}, $0{,}5h_{max}$, $0{,}25h_{max}$ oraz $0{,}125h_{max}$.

Tabela 6.1. Wyniki rozwiązania równania testowego jawną metodą Eulera

	$h = h_{max}$		$h = \frac{h_{max}}{2}$		$h = \frac{h_{max}}{4}$		$h = \frac{h_{max}}{8}$		
t_i	x	ε [%]	x	ε [%]	x	ε [%]	x	ε [%]	x_{dokl}
0,00	1	0	1	0	1	0	1	0	1
0,25	–	–	–	–	–	–	0,7500	3,7	0,77880
0,50	–	–	–	–	0,5	17,6	0,56250	7,3	0,60653
0,75	–	–	–	–	–	–	0,42188	10,7	0,47237
1,00	–	–	0	100	0,25	32,0	0,31641	14,0	0,36788
1,25	–	–	–	–	–	–	0,23730	17,2	0,28650
1,50	–	–	–	–	0,125	44,0	0,17798	20,2	0,22313
1,75	–	–	–	–	–	–	0,13348	23,2	0,17377
2,00	−1	838,9	0	100	0,0625	53,8	0,10011	26,0	0,13534
2,25	–	–	–	–	–	–	0,07509	28,8	0,10540
2,50	–	–	–	–	0,03125	61,9	0,05631	31,4	0,08209
2,75	–	–	–	–	–	–	0,04224	33,9	0,06393
3,00	–	–	0	100	0,01563	68,6	0,03168	36,4	0,04979

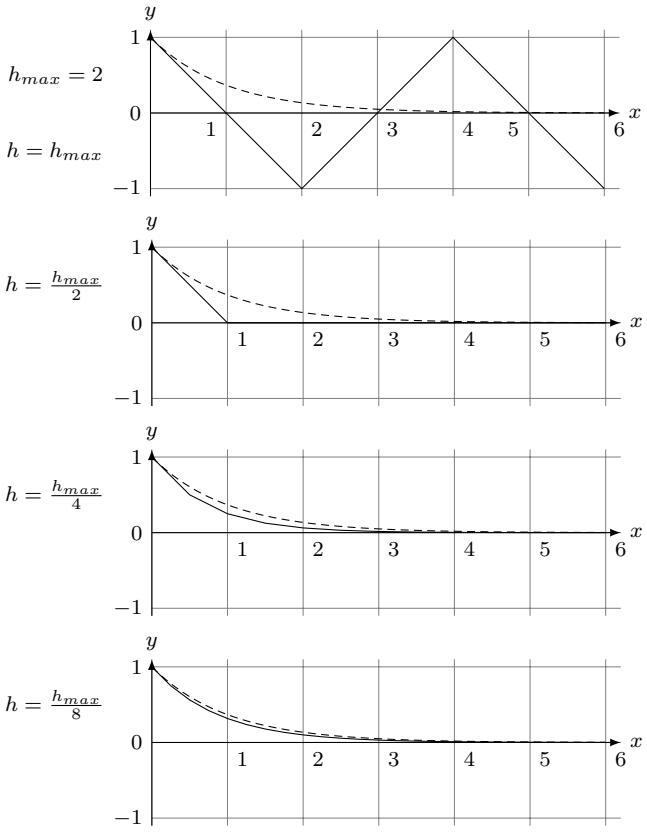
Rysunek 6.4 przedstawia rozwiązania dla poszczególnych długości kroku całkowania. Dobrze widoczna jest niestabilność rozwiązania, gdy długość kroku wynosi $h = 2$. W przypadku rozwiązywanego równania taka długość kroku stanowi granicę, od której jawna metoda Eulera traci stabilność. Mniejsze długości kroku dają coraz mniejszy błąd rozwiązania równania. Tabela 6.1, obejmująca wartości $t \in \langle 0; 3\rangle$ pozwala lepiej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

się zorientować w wielkości popełnianego błędu. Rozwiązanie dokładne analizowanego równania różniczkowego ma postać

$$x = e^{-t}$$

zaś wielkość błędu względnego przy najdokładniejszym kroku wynosi ponad 35%. Te wyniki wskazują, że nie należy traktować tej metody jako praktycznego narzędzia do rozwiązywania problemów inżynierskich.



Rysunek 6.4. Wyniki rozwiązania równania testowego jawną metodą Eulera z różnymi długościami kroku całkowania

Zasadniczą zaletą metody Eulera jest to, że wymaga bardzo niewielu działań algebraicznych, czyli można szybko wyznaczyć poszukiwane rozwiązanie. Ta właśnie cecha powoduje, że pomimo znacznie ograniczonego obszaru stabilności, znacznej niedokładności i bardzo prostej interpolacji rozwiązania, metoda Eulera i to w postaci jawnej jest stosowana w tych zagadnieniach, gdzie złożoność problemu, oznaczająca najczęściej jego wielkość, wymaga naprawdę szybkiej metody. Zaś rozwiązanie jest wystarczająco gładkie. Jeżeli występują duże różnice w stałych czasowych analizowanych

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

układów, wymagania dotyczące kroku czasowego mogą być tak ostre, że wykorzystanie tej metody do rozwiązania równania staje się nieefektywne.

Przykład obliczeniowy zastosowania niejawnej metody Eulera

Sformułowanie problemu. Niejawną metodą Eulera rozwiązać równanie różniczkowe pierwszego rzędu z warunkiem początkowym

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -x, \quad x(t_0) = 1$$

Przy rozwiązaniu użyć różnych długości kroku. Graniczna wielkość kroku całkowania wynosi $h_{max} = 2$. Równanie rozwiązać w przedziale $t \in \langle 0, 6\rangle$. Określić błąd względny rozwiązania.

Rozwiązanie. Wstawiając do zależności (6.10) prawą stronę rozwiązywanego równania, można wykonać następujące przekształcenia

$$x_{i+1} = x_i + h(-x_{i+1})$$

$$x_{i+1}(1 + h) = x_i$$

i uzyskać równanie rekurencyjne, pozwalające wyznaczać kolejne wartości rozwiązania równania niejawną metodą Eulera

$$x_{i+1} = \frac{x_i}{1 + h}$$

Część wartości uzyskanych przy rozwiązaniu zamieszczono w tabeli 6.2.

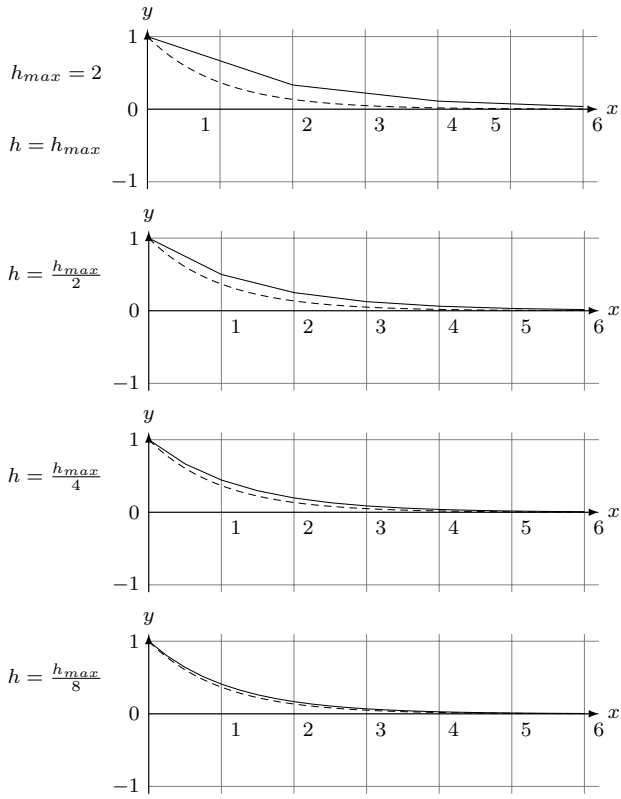
Tabela 6.2. Wyniki rozwiązania równania testowego niejawną metodą Eulera

	$h = h_{max}$		$h = \frac{h_{max}}{2}$		$h = \frac{h_{max}}{4}$		$h = \frac{h_{max}}{8}$		
t_i	x	ε [%]	x	ε [%]	x	ε [%]	x	ε [%]	x_{dokl}
0,00	1	0	1	0	1	0	1	0	1
0,25	–	–	–	–	–	–	0,8000	2,7	0,77880
0,50	–	–	–	–	0,66667	9,9	0,6400	5,5	0,60653
0,75	–	–	–	–	–	–	0,51200	8,4	0,47237
1,00	–	–	0,5	35,9	0,44444	20,8	0,40960	11,3	0,36788
1,25	–	–	–	–	–	–	0,32768	14,4	0,28650

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 6.2. cd.

1,50	–	–	–	–	0,29630	32,8	0,26214	17,5	0,22313
1,75	–	–	–	–	–	–	0,20972	20,7	0,17377
2,00	0,33	146	0,25	84,7	0,19753	46,0	0,16777	24,0	0,13534
2,25	–	–	–	–	–	–	0,13422	27,3	0,10540
2,50	–	–	–	–	0,13169	60,4	0,10737	30,8	0,08209
2,75	–	–	–	–	–	–	0,08590	34,4	0,06393
3,00	–	–	0,125	151	0,08779	76,3	0,06872	38,0	0,04979



Rysunek 6.5. Wyniki rozwiązania równania testowego niejawną metodą Eulera z różnymi długościami kroku całkowania

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 6.2 przedstawia wyznaczone wartości rozwiązania i wartości błędu dla długości kroku całkowania h równych kolejno h_{max}, $0{,}5h_{max}$, $0{,}25h_{max}$ oraz $0{,}125h_{max}$. Powyższa tabela 6.2 oraz rysunek 6.5 pokazują, że niejawna metoda Eulera jest stabilna przy tej samej długości kroku, przy której metoda jawna stawała się niestabilna. Jednak stabilność nie oznacza dokładności. Wielkości błędu przy najmniejszym kroku $h = 0{,}25$ są podobne. Zatem pomimo stabilności nie należy jej rozważać jako metody nadającej się do rozwiązywania problemów wymagających dokładności w praktyce inżynierskiej. Upatrywać jej zastosowania należy jedynie w specjalnych przypadkach.

Jeszcze krótko o błędach. Uważny czytelnik zauważy, iż chociaż teoria mówi, że metoda Eulera, w wariancie jawnym i niejawnym, jest rzędu pierwszego, co oznacza błąd proporcjonalny do h^2, to jednak przy analizie tabel 6.1 oraz 6.2 nie daje się łatwo tej zależności dostrzec. Zatem, jak to jest z tymi błędami?

Odpowiedź jest w miarę prosta. Otóż o błędach lokalnych mówi się wtedy, gdy punkt startowy dla danego kroku rozwiązania jest dokładnym rozwiązaniem równania różniczkowego. Wtedy różnica dokładnego i przybliżonego rozwiązania w następnym punkcie daje błąd lokalny. By dostrzec tę zależność, należy wybrać dwa kroki, każdy zaczynający się od wartości dokładnej.

Na przykład w przypadku niejawnym rozwiązanie dla $t = 1$ przy kroku $h = 1$ wynosi $x = 0{,}5$ z błędem prawie 36%. Gdy zaś rozpocznie się obliczenia od wartości $t = 0{,}5$ i $x = 0{,}60653$ (dokładnej!), to wynik dla $t = 1$ wyniesie $x(1) = 0{,}40435$, co daje błąd względny około 9,91%. Dobrze widać zależność kwadratową: dwa razy mniejszy krok, to cztery razy mniejszy błąd. Natomiast błędy w omawianych tabelach są błędami globalnymi, gdy obejmują więcej niż jeden krok, trudno wobec tego porównywać wynik z jednego kroku i kilku kroków.

6.2.3. Metoda Rungego–Kutty rzędu drugiego

Przez niektórych autorów jest uważana za metodę predyktor-korektor [42]. Wymaga wyznaczenia dwóch wartości funkcji w każdym kroku obliczeniowym. Wprowadza błąd lokalny rzędu $\mathcal{O}(h^3)$. Wyprowadzić ją można, wykorzystując rozwinięcie rozwiązania w szereg Taylora, jak to pokazano na przykład w pozycji [23], lub wykorzystując interpretację geometryczną, jak pokazano w [30]. Tutaj zostanie tylko podana ogólna zależność opisująca metodę. Równanie różniczkowe

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y)$$

można rozwiązać, korzystając ze wzorów (6.13). Te zależności w pozycji [33] zwane są metodą Eulera–Cauchy'ego, inni autorzy nazywają ją po prostu metodą Rungego–Kutty rzędu drugiego.

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + h, y_i + k_1) \\ y_{i+1} &= y_i + \tfrac{1}{2}(k_1 + k_2) \end{aligned} \tag{6.13}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Ponieważ w metodach Rungego–Kutty pewne współczynniki można dobierać dowolnie, ogólną postać metod Rungego–Kutty drugiego rzędu można zapisać jako

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + \kappa h, y_i + \kappa k_1) \\ y_{i+1} &= y_i + (1-\alpha)k_1 + \alpha k_2 \end{aligned}$$

Dodatnie parametry α i κ muszą spełniać zależność $2\alpha\kappa = 1$. W zależności od wartości parametrów α i κ uzyskuje się różne metody, wszystkie rzędu drugiego. Różnią się jedynie stałą błędu.

Lp.	α	κ	Metoda
1.	$\frac{1}{2}$	1	ulepszona metoda Eulera, lub standardowa Rungego–Kutty II rzędu (niektórzy autorzy nazywają ją metodą Heuna)
2.	1	$\frac{1}{2}$	metoda punktu środkowego
3.	$\frac{3}{4}$	$\frac{2}{3}$	metoda Heuna

Metoda Heuna daje najmniejszy błąd numeryczny, natomiast każda z wymienionych metod wprowadza w każdym kroku błąd proporcjonalny do h^3.

Ponieważ autorzy różnie nazywają poszczególne odmiany tej metody, trudno jest, bez szczegółowych badań, z całą pewnością uznać nazwy poszczególnych odmian. Z tego też powodu zamieszczone określenia należy traktować czysto informacyjnie i raczej jako ciekawostki niż obowiązujący standard nazewnictwa.

6.2.4. Przykład obliczeniowy

Sformułowanie problemu. Metodą Rungego–Kutty II rzędu rozwiązać równanie różniczkowe drugiego rzędu o stałych współczynnikach, opisujące ruch masy zawieszonej sprężyście z tłumieniem. Układ został przedstawiony na rysunku 6.6. Przyjąć, że masa $m = 5$ kg, współczynnik tłumienia $b = 1$ Ns/m, współczynnik sprężystości $k = 8$ N/m. Przyjąć, że warunki początkowe przy $t = 0$ wynoszą $x = 0{,}1$ m, $\dot{x} = 0$ m/s. Obliczenia wykonać z krokiem $h = 0{,}1$ s do chwili $t = 2$ s.

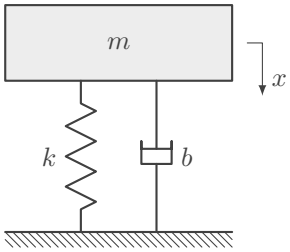
Dynamiczne równanie ruchu masy może być zapisane w następującej postaci

$$m\ddot{x} + b\dot{x} + kx = 0 \tag{6.14}$$

Zależność tę można zapisać w postaci układu równań pierwszego rzędu

$$\begin{bmatrix} \dot{z_1} \\ \dot{z_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 6.6. Model fizyczny analizowanego układu

Po wprowadzeniu oznaczeń macierzy przedstawionych w równaniu (6.8), otrzymuje się następujące macierzowe równanie różniczkowe pierwszego rzędu oraz warunek początkowy

$$\frac{\mathrm{d}\mathbb{z}}{\mathrm{d}t} = \mathbb{A}\mathbb{z}, \quad \mathbb{z}_0 = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_0 = \begin{bmatrix} 0{,}1 \\ 0 \end{bmatrix}$$

Uzyskane równanie łatwo można rozwiązać, wykorzystując wzory (6.13). Dla wyjaśnienia ewentualnych wątpliwości zostaną wykonane dwa kroki rozwiązania równania różniczkowego metodą Rungego-Kutty rzędu drugiego.

Krok 1. $i = 0$

$$k_1 = h\mathbb{A}\mathbb{z}_0 = h \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_0 = 0{,}1 \begin{bmatrix} 0 & 1 \\ -\frac{8}{5} & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} 0{,}1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -0{,}016 \end{bmatrix}$$

$$k_2 = h\mathbb{A}\left(\mathbb{z}_0 + k_1\right) = h \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_0 + k_1 \right) =$$

$$= 0{,}1 \begin{bmatrix} 0 & 1 \\ -\frac{8}{5} & -\frac{1}{5} \end{bmatrix} \left(\begin{bmatrix} 0{,}1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -0{,}016 \end{bmatrix} \right) = \begin{bmatrix} -0{,}0016 \\ -0{,}01568 \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_1 = \mathbb{z}_0 + \frac{1}{2}\left(k_1 + k_2\right) = \begin{bmatrix} 0{,}1 \\ 0 \end{bmatrix} + \frac{1}{2} \left(\begin{bmatrix} 0 \\ -0{,}016 \end{bmatrix} + \begin{bmatrix} -0{,}0016 \\ -0{,}01568 \end{bmatrix} \right) =$$

$$= \begin{bmatrix} 0{,}0992 \\ -0{,}01584 \end{bmatrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Punktem startowym dla kroku drugiego jest obliczony pierwszy punkt rozwiązania.

Krok 2. $i = 1$

$$k_1 = h\mathbb{A}\mathbb{z}_1 = h \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_1 = 0{,}1 \begin{bmatrix} 0 & 1 \\ -\frac{8}{5} & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} 0{,}0992 \\ -0{,}01584 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}001584 \\ -0{,}0155552 \end{bmatrix}$$

$$k_2 = h\mathbb{A}\left(\mathbb{z}_1 + k_1\right) = h \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix} \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_1 + k_1 \right) =$$

$$= 0{,}1 \begin{bmatrix} 0 & 1 \\ -\frac{8}{5} & -\frac{1}{5} \end{bmatrix} \left(\begin{bmatrix} 0{,}0992 \\ -0{,}01584 \end{bmatrix} + \begin{bmatrix} -0{,}001584 \\ -0{,}0155552 \end{bmatrix} \right) = \begin{bmatrix} -0{,}0031395 \\ -0{,}0149907 \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}_2 = \mathbb{z}_1 + \frac{1}{2}\left(k_1 + k_2\right) = \begin{bmatrix} 0{,}0992 \\ -0{,}01584 \end{bmatrix} +$$

$$+ \frac{1}{2} \left(\begin{bmatrix} -0{,}001584 \\ -0{,}0155552 \end{bmatrix} + \begin{bmatrix} -0{,}0031395 \\ -0{,}0149907 \end{bmatrix} \right) = \begin{bmatrix} 0{,}0968382 \\ -0{,}0311129 \end{bmatrix}$$

Wyniki dalszych obliczeń, aż do $t = 1$ s, zostały zebrane w tabeli 6.3. Natomiast rysunek 6.7 przedstawia przebieg przemieszczeń i prędkości masy m w zależności od czasu dla całego analizowanego zakresu.

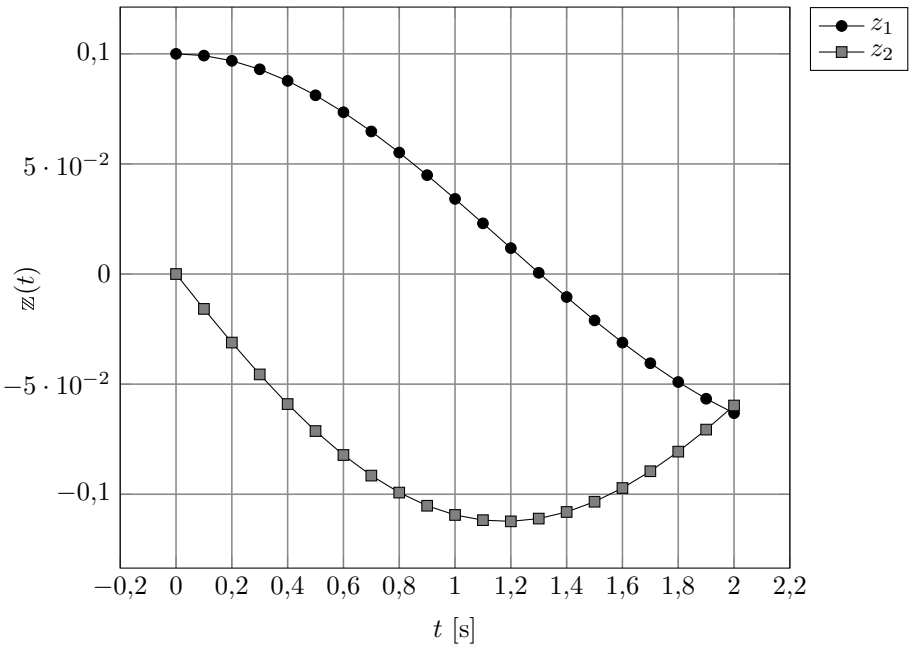
Tabela 6.3. Punkty $(1 - 10)$ rozwiązania równania (6.14) metodą Rungego–Kutty II rzędu

i	t_i	k_1^T		k_2^T		$\mathbb{z}_i$	
						z_1	z_2
0	0,0	—	—	—	—	0,10000	0,00000
1	0,1	0,00000	−0,01600	−0,00160	−0,01568	0,09920	−0,01584

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 6.3. cd.

2	0,2	−0,00158	−0,01556	−0,00314	−0,01499	0,09684	−0,03111
3	0,3	−0,00311	−0,01487	−0,00460	−0,01408	0,09298	−0,04559
4	0,4	−0,00456	−0,01397	−0,00596	−0,01296	0,08773	−0,05905
5	0,5	−0,00590	−0,01286	−0,00719	−0,01165	0,08118	−0,07130
6	0,6	−0,00713	−0,01156	−0,00829	−0,01019	0,07347	−0,08218
7	0,7	−0,00822	−0,01011	−0,00923	−0,00859	0,06475	−0,09153
8	0,8	−0,00915	−0,00853	−0,01001	−0,00689	0,05517	−0,09924
9	0,9	−0,00992	−0,00684	−0,01061	−0,00512	0,04490	−0,10522
10	1,0	−0,01052	−0,00508	−0,01103	−0,00329	0,03412	−0,10941

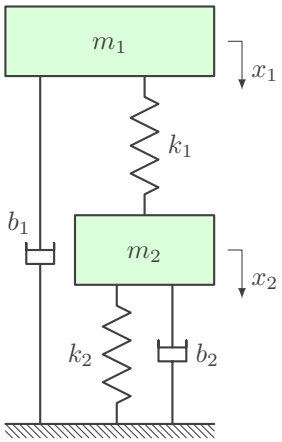


Rysunek 6.7. Czasowy przebieg rozwiązania równania (6.14)

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.2.5. Przykład

Sformułowanie problemu. Zapisać w postaci układu równań pierwszego rzędu układ dwóch równań różniczkowych zwyczajnych drugiego rzędu, opisujących zachowanie układu (rysunek 6.8) składającego się z dwóch mas zawieszonych sprężyście z tłumieniem bez zewnętrznych sił wymuszających.



Rysunek 6.8. Model fizyczny analizowanego układu

Rozwiązanie. Bez wnikania w zawiłości wyprowadzenia można zapisać równania dynamiczne ruchu układu, które mają postać następującą

$$\begin{cases} m_1\ddot{x}_1 + k_1(x_1 - x_2) + b_1\dot{x}_1 = 0 \\ m_2\ddot{x}_2 - k_1(x_1 - x_2) + k_2x_2 + b_2\dot{x}_2 = 0 \end{cases}$$

Następnie, zgodnie ze sposobem postępowania przedstawionym w podrozdziale 6.2.1, należy przekształcić ten układ równań do postaci

$$\begin{cases} \ddot{x}_1 = -\frac{k_1}{m_1}(x_1 - x_2) - \frac{b_1}{m_1}\dot{x}_1 \\ \ddot{x}_2 = \frac{k_1}{m_2}(x_1 - x_2) - \frac{k_2}{m_2}x_2 - \frac{b_2}{m_2}\dot{x}_2 \end{cases}$$

Zastosowane zostaną następujące podstawienia zmiennych pomocniczych z_i

$$\begin{cases} z_1 = x_1 \\ z_2 = x_2 \\ z_3 = \dot{z}_1 \\ z_4 = \dot{z}_2 \end{cases}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po wprowadzeniu zmiennych pomocniczych należy wyrazić wszystkie przemieszczenia i prędkości za ich pomocą. W równaniach nie mogą się pojawić zmienne x_1, x_2 oraz ich pochodne. Zatem analizowany układ równań przyjmuje postać następującą

$$\begin{cases} \dot{z}_1 = z_3 \\ \dot{z}_2 = z_4 \\ \dot{z}_3 = -\frac{k_1}{m_1} z_1 + \frac{k_1}{m_1} z_2 - \frac{b_1}{m_1} z_3 \\ \dot{z}_4 = \frac{k_1}{m_2} z_1 - \frac{k_1}{m_2} z_2 - \frac{k_2}{m_2} z_2 - \frac{b_2}{m_2} z_4 \end{cases}$$

Tak przekształcony układ równań różniczkowych pierwszego rzędu można zapisać w postaci macierzowej w następujący sposób

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k_1}{m_1} & \frac{k_1}{m_1} & -\frac{b_1}{m_1} & 0 \\ \frac{k_1}{m_2} & -\frac{k_1+k_2}{m_2} & 0 & -\frac{b_2}{m_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix}$$

Po wprowadzeniu oznaczeń macierzy przedstawionych w równaniu (6.8), otrzymuje się następujące macierzowe równanie różniczkowe pierwszego rzędu

$$\frac{\mathrm{d}\mathbb{z}}{\mathrm{d}t} = \mathbb{A}\mathbb{z}$$

Aby można było rozwiązać uzyskane równanie różniczkowe, konieczne jest określenie warunków początkowych i kroku całkowania. Następnie można przystąpić do jego rozwiązania dowolną z omawianych metod rozwiązywania równań różniczkowych zwyczajnych rzędu pierwszego.

6.2.6. Metoda Rungego-Kutty rzędu czwartego

Historycznie najstarsza z metod Rungego–Kutty. Służy do rozwiązywania równań różniczkowych pierwszego rzędu w postaci

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y)$$

z zadanym warunkiem początkowym $y(x_0) = y_0$. Gdy prawa strona równania zależy jedynie od zmiennej niezależnej x, można to równanie rozwiązać tradycyjnymi metodami całkowania numerycznego. Sytuacja się komplikuje, gdy po prawej stronie

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

występuje zależność od zmiennej y. Takie równanie można rozwiązać, korzystając z następujących wzorów zwanych metodą Rungego–Kutty rzędu czwartego

$$\begin{cases} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}) \\ k_3 &= hf(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}) \\ k_4 &= hf(x_i + h, y_i + k_3) \\ y_{i+1} &= y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \end{cases}$$

Metoda ta jest też znana po prostu jako metoda Rungego–Kutty. Przy wyprowadzaniu metod Rungego–Kutty, omówionym na przykład w publikacjach [23, 29, 38], uzyskuje się układ zależnych liniowych równań algebraicznych. To powoduje, że można wybrać różne układy współczynników i punktów, w których są wyznaczane dodatkowe wartości funkcji. Dzięki temu istnieje wiele różnych odmian metod Rungego–Kutty danego rzędu. Niektóre z nich mają swoje indywidualne nazwy, a różnią się jedynie wartością stałej błędu.

Podobnie jak w innych różnicowych metodach jednokrokowych wyznaczenie następnej wartości rozwiązania $y_{i+1} = y(x_{i+1})$ odbywa się na podstawie wartości bieżącej y_i, do której dodaje się ważoną średnią współczynników. Liczba współczynników oznacza liczbę etapów metody, w tym przypadku są to cztery etapy. Współczynniki k_j mają jednakową postać. Każdy z nich wylicza się, mnożąc długość kroku całkowania przez tangens kąta nachylenia stycznej do rozwiązania na początku, w połowie i na końcu przedziału całkowania. Ów tangens kąta nachylenia stycznej do rozwiązania jest dany prawą stroną rozwiązywanego równania. Opisane współczynniki stanowią przyrosty i można je interpretować w następujący sposób:

— k_1 jest przyrostem określonym na podstawie nachylenia stycznej na początku przedziału, tak jak w metodzie Eulera;
— k_2 jest przyrostem określonym w połowie przedziału dla wartości y_i, do której dodano połowę przyrostu z metody Eulera;
— k_3 stanowi przyrost także określany w połowie długości kroku, ale do y_i dodaje się połowę przyrostu k_2 określonego na podstawie nachylenia stycznej w połowie przedziału;
— k_4 to przyrost określany na końcu przedziału całkowania, gdzie do y_i dodaje się cały przyrost k_3.

Wagi ustalone są w taki sposób, by uśrednianie odbywało się na końcu przedziału całkowania, czyli te przyrosty, które były wyznaczane w połowie przedziału, są mnożone przez wagę równą dwa.

Jak wspomniano na początku tego rozdziału, metoda Rungego–Kutty rzędu czwartego daje błąd lokalny rzędu $\mathcal{O}(h^5)$, zaś globalny rzędu $\mathcal{O}(h^4)$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.2.7. Ogólne informacje o metodach Rungego–Kutty

W niniejszym skrypcie główny nacisk położony został na jawne metody Rungego–Kutty, istnieją też niejawne ich wersje, które oczywiście mają znacznie większy obszar stabilności. To powoduje, że są używane do rozwiązywania zadań sztywnych. Jak już kilkukrotnie stwierdzono, jawne metody Rungego–Kutty zawodzą przy próbach rozwiązywania takich zadań. Kilka niejawnych metod Rungego–Kutty zostanie przedstawionych w dalszej części rozdziału.

Mówi się, że metoda Rungego–Kutty jest m-etapowa, jeżeli do obliczenia następnej wartości rozwiązania równania różniczkowego konieczne jest obliczenie m razy wartości funkcji. To właśnie liczba etapów metody determinuje możliwy do uzyskania rząd metody. I tak, przedstawiona do tej pory metoda rzędu drugiego była dwuetapowa, natomiast metoda rzędu czwartego – czteroetapowa.

Badacze, szczególnie John Charles Butcher, udowodnili że, dla metod jawnych Rungego–Kutty, liczba etapów m i rząd metody p możliwy do uzyskania przy danej liczbie etapów są powiązane. Związek ten został przedstawiony w tabeli 6.4 W tabelach 6.4 oraz 6.5 dla czytelności do zapisu rzędu metody zastosowano cyfry arabskie, zamiast jak przyjęto w pozostałych miejscach książki cyfry rzymskie.

Tabela 6.4. Możliwy do osiągnięcia rząd metody dla danej liczby etapów w jawnych metodach Rungego–Kutty

	$p = m$				$p = m - 1$			$p = m - 2$	
Liczba etapów m	1	2	3	4	5	6	7	8	9
Rząd metody p	1	2	3	4	4	5	6	6	7

Inną kłopotliwą właściwością metod Rungego–Kutty jest to, że jeżeli metoda jest rzędu p niższego niż cztery dla pojedynczego równania, to jest ona też tego samego rzędu dla układu równań. Natomiast gdy metoda jest rzędu wyższego niż cztery dla pojedynczego równania, to nie musi tak być dla układu równań [29]. Między innymi, to ta cecha i złożoność omawianych metod rozwiązywania równań różniczkowych zwyczajnych odpowiadają za to, że dosyć rzadko stosuje się metody rzędów wyższych niż cztery.

Stabilność metod Rungego–Kutty stanowi ich cenną zaletę. Okazuje się, że wraz ze wzrostem rzędu metody, rośnie też jej obszar stabilności. Co powoduje, że mimo większego kosztu obliczeniowego, opłaca się stosować metody wyższych rzędów ze względu na możliwość używania dłuższych kroków całkowania. Dla metod Rungego–Kutty wielkość obszaru stabilności przedstawia tabela 6.5. Porównując obszary stabilności jawnych metod Rungego–Kutty z przedstawionymi nieco dalej obszarami stabilności metod wielokrokowych, można zauważyć, że niejawne metody wielokroko-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

we do trzeciego rzędu mają większy obszar stabilności, zaś metoda Rungego–Kutty rzędu czwartego ma już większy obszar stabilności.

Tabela 6.5. Obszary stabilności metod Rungego–Kutty

Rząd	$h\lambda \in$
1	$(-2, 0)$
2	$(-2, 0)$
3	$(-2.51, 0)$
4	$(-2.78, 0)$

Wielokrokowe metody jawne przegrywają pod tym względem z metodami Rungego–Kutty. Niejawne metody Rungego–Kutty mają jeszcze większe obszary stabilności, chociaż są znacznie trudniejsze w zastosowaniu i wymagają znacznie większej liczby działań arytmetycznych niż ich jawne odpowiedniki.

6.2.8. Tablice Butchera

W 1964 roku John Charles Butcher napisał artykuł [8], w którym metody Rungego–Kutty przedstawił w postaci tablicowej. Dzięki temu łatwiej jest je zapisywać. Zanim przedstawiony zostanie sposób tworzenia takiej tablicy, krótkie wprowadzenie.

W 1901 roku Martin Wilhelm Kutta uogólnił wcześniejsze prace Carla Rungego i Karla Heuna, wprowadzając określenie, które można ująć w następujący sposób. Gdy s jest liczbą etapów, zaś a_{21}, a_{31}, a_{32}, ..., a_{s1}, a_{s2},..., $a_{s,s-1}$, b_1, b_2, ..., b_s, c_2,..., c_s są rzeczywistymi współczynnikami, to zależność

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + c_2 h, y_i + a_{21} k_1) \\ k_3 &= hf(x_i + c_3 h, y_i + a_{31} k_1 + a_{32} k_2) \\ &\dots \\ k_s &= hf(x_i + c_s h, y_i + a_{s1} k_1 + \ldots + a_{s,s-1} k_{s-1}) \\ y_{i+1} &= y_i + b_1 k_1 + \ldots + b_s k_s \end{aligned} \tag{6.15}$$

nazywa się s-etapową metodą Rungego-Kutty. Zazwyczaj współczynniki c_i spełniają warunki

$$c_2 = a_{21}, \quad c_3 = a_{31} + a_{32}, \quad \cdots \quad c_s = a_{s1} + \ldots + a_{s,s-1} \tag{6.16}$$

lub w skrócie

$$c_i = \sum_{j=1}^{i-1} a_{ij}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zgodnie z koncepcją Butchera, metody Rungego–Kutty, zapisane zależnościami (6.15), można zwięźle zapisać w postaci tablicy (6.17).

$$\begin{array}{c|cccccc} 0 & & & & & & \\ c_2 & a_{21} & & & & & \\ c_3 & a_{31} & a_{32} & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} & & \\ \hline & b_1 & b_2 & \dots & b_{s-1} & b_s \end{array} \tag{6.17}$$

Tablice w metodach jawnych są trójkątne, zaś w metodach niejawnych – kwadratowe. Przez to wymagają w każdym kroku rozwiązania układu algebraicznych równań liniowych. Wcześniej wymienione metody można zapisać za pomocą tej notacji w następujący sposób.

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Metoda Rungego–Kutty rzędu IV w postaci klasycznej

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Metoda Rungego–Kutty rzędu II

Jak widać, tablice Butchera to bardzo efektywny sposób zapisu metod Rungego–Kutty. Dzięki nim można bardzo zwięźle ująć rozbudowane zależności, które występują szczególnie w metodach wyższych rzędów.

6.2.9. Metody Rungego–Kutty z estymacją błędu

Autorzy Erwin Fehlberg, R. H. Merson, Jim Verner Prince, John Dormand i inni konstruowali metody Rungego–Kutty w ten sposób, że dla tych samych współczynników c_i oraz a_{ij} wyznaczali dwa różne zestawy współczynników b_i, uzyskując w ten sposób połączone metody rzędów różniących się o jeden [9]. Takie zestawy metod nazywane są po angielsku *embedded*, co można przetłumaczyć jako osadzone. Poniżej zostanie przedstawiona jedna taka metoda oraz sposób szacowania błędu. Metodę

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rungego–Kutty–Vernera rzędu VI, z kontrolą błędu rzędu V, co zapisuje się jako rzędu VI(V), można opisać tablicą Butchera (6.18).

$$
\begin{array}{c|cccccccc}
0 & & & & & & & & \\
\frac{1}{6} & \frac{1}{6} & & & & & & & \\
\frac{4}{15} & \frac{4}{75} & \frac{16}{75} & & & & & & \\
\frac{2}{3} & \frac{5}{6} & -\frac{8}{3} & \frac{5}{2} & & & & & \\
\frac{5}{6} & -\frac{165}{64} & \frac{55}{6} & -\frac{425}{64} & \frac{85}{96} & & & & \\
1 & \frac{12}{5} & -8 & \frac{4015}{612} & -\frac{11}{36} & \frac{88}{255} & & & \\
\frac{1}{15} & -\frac{8263}{15000} & \frac{124}{75} & -\frac{643}{680} & -\frac{81}{250} & \frac{2484}{10625} & 0 & & \\
1 & \frac{3501}{1720} & -\frac{300}{43} & \frac{297275}{52632} & -\frac{319}{2322} & \frac{24068}{84065} & 0 & \frac{3850}{26703} & \\
\hline
y_1 & \frac{3}{40} & 0 & \frac{875}{2244} & \frac{23}{72} & \frac{264}{1955} & 0 & \frac{125}{11592} & \frac{43}{616} \\
\hat{y}_1 & \frac{13}{160} & 0 & \frac{2375}{5984} & \frac{5}{16} & \frac{12}{85} & \frac{3}{44} & 0 & 0
\end{array}
\tag{6.18}
$$

Współczynniki w dwóch ostatnich rzędach dotyczące y_1 oraz $\hat{y}_1$ służą do wyznaczania rozwiązania oraz określania wielkości błędu. Lokalny błąd rozwiązania w przybliżeniu wynosi

$$\Delta = y_1 - \hat{y}_1 = \sum_i (b_i - \hat{b}_i) k_i$$

Jak widać, szacowanie błędu jest niezwykle eleganckie i wydajne obliczeniowo. Nie wymaga obliczania wartości funkcji, a jedynie kombinacji liniowej różnicy wag oraz współczynników k_i.

Jedną z ostatnich propozycji nowej pary metod z rodziny Rungego–Kutty można znaleźć w artykule [47] opisującym metodę z obniżoną liczbą założeń nakładanych na równania służące do wyznaczania współczynników. Efektem jest obniżenie wielkości błędu, w porównaniu do standardowej pary metod Dormanda i Prince'a.

Czytelnik zainteresowany innymi parami metod może sięgnąć do obfitej literatury z tego zakresu, na przykład publikacje [18, 23] czy inne.

6.2.10. Niejawne metody Rungego–Kutty

Niejawne metody Rungego–Kutty opisują tablice Butchera, w których współczynniki a_{ij} tworzą macierz kwadratową, a nie trójkątną dolną jak jest przy metodach

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

jawnych. Zatem ogólna postać tych metod przedstawiona została zależnością (6.19).

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \tag{6.19}$$

Zastosowanie tych metod będzie wymagało w każdym kroku rozwiązania układu równań. Nie wnikając w zawiłości wyprowadzenia, poniżej zostaną przedstawione dwie metody niejawne z rodziny metod Rungego–Kutty [8, 19].

Pierwsza z nich, zwana metodą Gaussa rzędu drugiego, jest wyrażana następującą tablicą Butchera.

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Z kolei metoda rzędu czwartego może zostać wyrażona tablicą przedstawioną poniżej.

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Natomiast metoda rzędu szóstego została przedstawiona poniżej.

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Wymienione metody bazujące na kwadraturach Gaussa są A-stabilne. Z tego powodu są zalecane do rozwiązywania równań, które ze względu na utratę stabilności wymuszają stosowanie niewielkich kroków całkowania przy stosowaniu metod jawnych. Takie równania nazywane są równaniami sztywnymi.

6.2.11. Metody wielokrokowe

Dotychczas omawiane metody należą do klasy metod jednokrokowych, co oznacza, że do wyznaczenia następnej wartości poszukiwanego rozwiązania jest wykorzystywana jedynie aktualna wartość rozwiązania. Jak można zauważyć, w metodach

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rungego–Kutty czasem korzysta się z pośrednich wartości funkcji, leżących między aktualnym i następnym punktem rozwiązania. Niemniej zawsze następna wartość rozwiązania jest obliczana na podstawie aktualnej wartości. Inną rodzinę metod stanowią metody wielokrokowe, w których następną wartość rozwiązania oblicza się na podstawie kilku poprzednio obliczonych wartości. Omawiane metody wynikają z przedstawienia rozwiązania równania różniczkowego pierwszego rzędu

$$\frac{\mathrm{d}x}{\mathrm{d}y} = f(x, y), \quad y(x_0) = y_0$$

w postaci następującej

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(x, y)\mathrm{d}x$$

Całkę po prawej stronie tego równania można obliczyć numerycznie, dokonując interpolacji wielomianowej funkcji podcałkowej $f(x, y)$, a następnie całkując uzyskany wielomian. Uzyskane w ten sposób metody można wyrazić wzorem jawnym

$$y_{i+1} = y_i + af(x_i, y_i) + bf(x_{i-1}, y_{i-1}) + cf(x_{i-2}, y_{i-2}) + \cdots \tag{6.20}$$

zwanym wzorem Adamsa–Bashfortha oraz niejawnym

$$y_{i+1} = y_i + af(x_{i+1}, y_{i+1}) + bf(x_i, y_i) + cf(x_{i-1}, y_{i-1}) + df(x_{i-2}, y_{i-2}) + \cdots \tag{6.21}$$

nazywanym wzorem Adamsa–Moultona. Te ogólne zależności opisują dwie liczne rodziny metod. Najprostszą metodą, będącą jednocześnie metodą pierwszego rzędu, jest omawiana wcześniej metoda Eulera. Warto też wspomnieć o innej szeroko stosowanej metodzie wielokrokowej.

6.2.12. Niejawna metoda trapezów

Jest to stabilna metoda drugiego rzędu, o najmniejszym błędzie w klasie metod nadających się do rozwiązywania zadań sztywnych [23]. Metoda trapezów opisana jest zależnością

$$y_{i+1} = y_i + \frac{h}{2}\left(f(x_{i+1}, y_{i+1}) + f(x_i, y_i)\right) \tag{6.22}$$

Chociaż jest to metoda jednokrokowa, to jednak niejawna metoda trapezów należy do rodziny metod wielokrokowych Adamsa–Moultona, co zostanie pokazane na końcu niniejszego rozdziału. Błąd lokalny popełniany w każdym kroku jest rzędu $\mathcal{O}(h^3)$, co oznacza, że metoda jest rzędu drugiego. Obszar stabilności niejawnej metody trapezów obejmuje całą lewą półpłaszczyznę [29]. Oznacza, że metoda jest A-stabilna. Dzięki A-stabilności można tę metodę stosować do rozwiązywania równań źle uwarunkowanych nazywanych też sztywnymi.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wzory Adamsa–Bashfortha

Są wzorami w ogólnej postaci (6.20). Szczegółowe zależności dla rzędów od 1. do 5. zamieszczono w tabeli 6.7. Ponieważ są to metody jawne, są wrażliwe na długość kroku całkowania. Dla dłuższych kroków możliwe są niestabilności. Obszary stabilności absolutnej metod Adamsa–Bashfortha (A-B) oraz Adamsa–Moultona (A-M) zostały zestawione w tabeli 6.6.

Tabela 6.6. Obszary stabilności metod Adamsa

Rząd	A-B	A-M
	$h\lambda \in$	
1	$(-2, 0)$	$(-\infty, 0)$
2	$(-1, 0)$	$(-6, 0)$
3	$(-\frac{6}{11}, 0)$	$(-3, 0)$
4	$(-\frac{3}{10}, 0)$	$(-\frac{90}{49}, 0)$

Widać, iż metody Adamsa–Moultona mają znacznie większy obszar stabilności, natomiast w obu przypadkach te obszary maleją ze wzrostem rzędu metody, co stanowi zachowanie przeciwne do tego obserwowanego przy metodach Rungego–Kutty.

Wzory Adamsa–Moultona

Pod nazwą wzorów Adamsa–Moultona znane są zależności w ogólnej postaci (6.21). Poniżej w tabeli 6.7 zestawiono wzory Adamsa–Bashfortha oraz, dla porównania, Adamsa–Moultona rzędów od 1. do 5. Są to metody niejawne (*implicit*), ponieważ po prawej stronie tych zależności znajduje się nieznana wartość funkcji $f(x_{i+1}, y_{i+1})$ w każdym kroku rozwiązania równania różniczkowego należy rozwiązać równanie nieliniowe, co prowadzi do znaczącego zwiększenia kosztu obliczeniowego. Strategie rozwiązywania równań nieliniowych są różne, niektórzy autorzy (publikacje [5, 23, 29]) zalecają wykonanie kilku iteracji metodą iteracji prostej, inni proponują metodę Newtona lub siecznych stosowane do uzyskania zadowalającej dokładności. Słabością metody iteracji prostej jest jej mocno ograniczony obszar stabilności, który powoduje, iż niekiedy nie daje się uzyskać zbieżności rozwiązania uzyskanego równania nieliniowego. Z kolei metoda Newtona jest znacznie bardziej złożona obliczeniowo.

Metody predyktor-korektor

Metody predyktor-korektor zwane też ekstrapolacyjno-interpolacyjnymi, są bardzo wydajnym połączeniem metod jawnych i niejawnych. W zasadzie jest to realizacja metody niejawnej z wykorzystaniem metody iteracji prostej do rozwiązania uzyskanego nieliniowego równania algebraicznego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 6.7. Jawne i niejawne metody Adamsa, przyjęto oznaczenia $f(x_i, y_i) = f_i$

Rząd	Wzory Adamsa–Bashfortha
1	$y_{i+1} = y_i + hf_i$
2	$y_{i+1} = y_i + \frac{h}{2}(3f_i - f_{i-1})$
3	$y_{i+1} = y_i + \frac{h}{12}(23f_i - 16f_{i-1} + 5f_{i-2})$
4	$y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3})$
5	$y_{i+1} = y_i + \frac{h}{720}(1901f_i - 2774f_{i-1} + 2616f_{i-2} - 1274f_{i-3} + 251f_{i-4})$
Rząd	**Wzory Adamsa–Moultona**
1	$y_{i+1} = y_i + hf_{i+1}$
2	$y_{i+1} = y_i + \frac{h}{2}(f_{i+1} + f_i)$
3	$y_{i+1} = y_i + \frac{h}{12}(5f_{i+1} + 8f_i - f_{i-1})$
4	$y_{i+1} = y_i + \frac{h}{24}(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2})$
5	$y_{i+1} = y_i + \frac{h}{720}(251f_{i+1} + 646f_i - 264f_{i-1} + 106f_{i-2} - 19f_{i-3})$

Wartość początkowa jest wyznaczana za pomocą metody jawnej, po czym przechodzi się do kolejnych iteracji metodą niejawną. W ogólnym przypadku metody predyktor-korektor składają się z dwóch metod:

— metody jawnej służącej do wyznaczenia przybliżonej wartości następnego punktu rozwiązania oraz
— metody niejawnej, pozwalającej skorygować poprzednio wyznaczoną wartość.

Wzory Adamsa–Bashfortha i Adamsa–Moultona stanowią dobrze dobraną parę znajdującą zastosowanie w metodach typu predyktor-korektor.

Przy ustalonym rzędzie metody, najpierw za pomocą wzoru Adamsa–Bashfortha realizuje się krok, wyznaczając wartość y_{i+1}. Skoro jest znany następny punkt rozwiązania, można skorzystać z wzoru Adamsa–Moultona (startując z punktu y_i) i poprawić wartość y_{i+1}. W ten sposób unika się jawnego rozwiązywania równań nieliniowych w każdym kroku rozwiązania równania różniczkowego, przy stosunkowo dużej dokładności rozwiązania. Aby zwiększyć dokładność, można powtórzyć etap korekcji (nawet kilkakrotnie), korzystając z uzyskanego poprzednio przybliżenia. W ten sposób wykorzystuje się metodę iteracji prostej do rozwiązania równania nieliniowego, które jest wynikiem działania metody niejawnej. Można oczywiście użyć innej metody rozwiązywania równań nieliniowych, znacznie szybciej zbieżnej niż metoda iteracji

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

prostej, co wiąże się ze zwiększeniem kosztu obliczeniowego. Predyktor służy jedynie do dostarczenia korektorowi dobrego przybliżenia następnego punktu rozwiązania. Gdy zostanie wykonanych odpowiednio dużo iteracji lub wystarczająco dokładnie rozwiąże się równanie nieliniowe, maleje wpływ predyktora na rozwiązanie i stabilność metody predyktor-korektor. Jeżeli jednak wykona się niewielką liczbę iteracji (zazwyczaj wykonuje się nie więcej niż trzy), to przy określaniu stabilności metody należy uwzględnić wpływ predyktora.

Metody Geara

Metody Geara są jeszcze innym przedstawicielem rodziny metod wielokrokowych [28]. Nazywane są też metodami różniczkowania wstecznego (*backward differentiation formula*). Są one opisane następującą ogólną zależnością

$$y_{i+1} = h\beta_0 f(x_{i+1}, y_{i+1}) + \sum_{k=1}^{K} \alpha_k y_{i+1-k} \tag{6.23}$$

Warto zauważyć, iż w odróżnieniu od omówionych powyżej metod Adamsa, w metodach Geara nie występuje wartość bieżąca funkcji $f(x_i, y_i)$ ani poprzednie jej wartości, a jedynie wartość rozwiązania y_i. W związku z tym są one metodami niejawnymi i wymagają w każdym kroku rozwiązania równania nieliniowego. Współczynniki metod Geara dla rzędów od 1. do 6. przedstawia tabela 6.8.

Tabela 6.8. Metody Geara rzędów od 1. do 6.

K	β_0	α_1	α_2	α_3	α_4	α_5	α_6
1	1	1					
2	$\frac{2}{3}$	$\frac{4}{3}$	$-\frac{1}{3}$				
3	$\frac{6}{11}$	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$			
4	$\frac{12}{25}$	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$-\frac{3}{25}$		
5	$\frac{60}{137}$	$\frac{300}{137}$	$-\frac{300}{137}$	$\frac{200}{137}$	$-\frac{75}{137}$	$\frac{12}{137}$	
6	$\frac{60}{147}$	$\frac{360}{147}$	$-\frac{450}{147}$	$\frac{400}{147}$	$-\frac{225}{147}$	$\frac{72}{147}$	$-\frac{10}{147}$

Metody rzędu większego niż szósty nie są używane ze względu na ich niestabilność. Zasadnicza zaletą tych metod jest to, że można nimi rozwiązywać tzw. równania sztywne, ze względu na ich duży obszar stabilności.

Metoda Newmarka

W odróżnieniu od dotychczas omawianych metod metoda Newmarka, a ściślej rzecz biorąc – metoda βNewmarka, służy do bezpośredniego rozwiązywania równań

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

różniczkowych zwyczajnych drugiego rzędu. Została opisana około 1959 roku przez Nathana M. Newmarka. Dla danego równania ruchu w postaci

$$m\ddot{x} + b\dot{x} + kx = F \tag{6.24}$$

wprowadza się dyskretyzację w dziedzinie czasu i dla chwili t_{i+i} można zapisać następująco

$$m\ddot{x}_{i+1} + b\dot{x}_{i+1} + kx_{i+1} = F_{i+1} \tag{6.25}$$

Zakładając odpowiedni przebieg zmian przyspieszenia, można dosyć ogólnie napisać następujące zależności na prędkość i przemieszczenie

$$\dot{x}_{i+1} = \dot{x}_i + [(1-\beta)\ddot{x}_i + \beta\ddot{x}_{i+1}]h \tag{6.26}$$

$$x_{i+1} = x_i + h\dot{x}_i + \left[\left(\frac{1}{2} - \alpha\right)\ddot{x}_i + \alpha\ddot{x}_{i+1}\right]h^2 \tag{6.27}$$

parametry α oraz β określają wpływ przyspieszenia na prędkość i przemieszczenie w kolejnych krokach, i tak wartości:

— $\alpha = 1/6, \beta = 1/2$ dają liniową zmianę przyspieszenia w każdym kroku, zaś
— $\alpha = 1/4, \beta = 1/2$ dają stałą wartość przyspieszenia w każdym kroku.

Więcej informacji na temat właściwości metody w zależności od parametrów α i β zawiera tabela 6.9.

W równaniach (6.26) oraz (6.27) kłopotliwe zdają się wartości przyspieszenia w następnej chwili czasowej $\ddot{x}_{i+1}$. Można je wyznaczyć z równania (6.27) i wstawić do równania (6.26), przez co uzyskuje się

$$\ddot{x}_{i+1} = \frac{1}{\alpha h^2}\left[x_{i+1} - x_i - h\dot{x}_i - h^2\left(\frac{1}{2} - \alpha\right)\ddot{x}_i\right] \tag{6.28}$$

$$\dot{x}_{i+1} = \frac{\beta}{\alpha h}(x_{i+1} - x_i) + \left(1 - \frac{\beta}{\alpha}\right)\dot{x}_i + h\left(1 - \frac{\beta}{2\alpha}\right)\ddot{x}_i \tag{6.29}$$

Po podstawieniu powyższych zależności do (6.25) uzyska się

$$\left[\frac{1}{\alpha h^2}m + b\frac{\beta}{\alpha h} + k\right]x_{i+1} = F_{i+1} + m\left[\frac{1}{\alpha h^2}x_i + \frac{1}{\alpha h}\dot{x}_i + \left(\frac{1}{2\alpha} - 1\right)\ddot{x}_i\right] + \\ + b\left[\frac{\beta}{\alpha h}x_i + \left(\frac{\beta}{\alpha} - 1\right)\dot{x}_i + h\left(\frac{\beta}{2\alpha} - 1\right)\ddot{x}_i\right] \tag{6.30}$$

Wykorzystując powyższą zależność, można wyznaczyć nieznane przemieszczenie x_{i+1}. Przy znanym przemieszczeniu w następnej chwili, można wyznaczyć prędkość i przyspieszenie w następnym kroku z zależności (6.28) i (6.29). Warto zapamiętać, iż dla parametrów $\beta \geqslant 0{,}5$ i $\alpha \geqslant 0{,}25(\beta + 0{,}5)^2$ metoda Newmarka jest bezwarunkowo stabilna. W wielu publikacjach można znaleźć wartości $\beta = 0{,}5$ oraz $\alpha = 0{,}25$ podawane jako zalecane dla analizy układów dynamicznych. Metoda jest szczególnie przydatna przy analizie dynamicznej z wykorzystaniem metody elementów skończonych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Algorytm 6.1. Metoda Newmarka

Require: h, x_0, $\dot{x}_0$, $\ddot{x}_0$, α, β, m, b, k, t_0, t_k

$a_0 \leftarrow \frac{1}{h^2\alpha}$
$a_2 \leftarrow \frac{1}{h\alpha}$
$a_1 \leftarrow \beta a_2$
$a_3 \leftarrow \frac{1}{2\alpha} - 1$
$a_4 \leftarrow \frac{\beta}{\alpha} - 1$
$a_5 \leftarrow \frac{h}{2}\left(\frac{\beta}{\alpha} - 2\right)$
$a_6 \leftarrow h(1-\beta)$
$a_7 \leftarrow h\beta$
$\hat{k} \leftarrow k + a_0 m + a_1 b$
$x(t_0) \leftarrow x_0$
$\dot{x}(t_0) = \dot{x}_0$
$\ddot{x}(t_0) = m^{-1}(F(t_0) - b\dot{x}(t_0) - kx(t_0))$
for $li = t_0$ to t_k **do**
  $F(li) \leftarrow F(li-1) + m(a_0 x(li) + a_2\dot{x}(li) + a_3\ddot{x}(li)) + b(a_1 x(li) + a_4\dot{x}(li) + a_5\ddot{x}(li))$
  $x(li+1) \leftarrow \hat{k}^{-1} Fn$
  $\ddot{x}(li+1) \leftarrow a_0(x(li+1) - x(li)) - a_2\dot{x}(li) - a_3\ddot{x}(li)$
  $\dot{x}(li+1) \leftarrow \dot{x}(li) + a_6\ddot{x}(li) + a_7\ddot{x}(li+1)$
end for

Tabela 6.9. Obszar stabilności, błąd bezwzględny amplitudy i błąd względny okresu w różnych wariantach metody Newmarka

Metoda	β	α	Granica stab. $h\omega$	Błąd ampl.	Błąd okresu $\frac{\Delta T}{T}$
Liniowe przyspieszenie	$\frac{1}{2}$	$\frac{1}{6}$	3,46	0	$\frac{h^2\omega^2}{24}$
Stałe średnie przyspieszenie	$\frac{1}{2}$	$\frac{1}{4}$	∞	0	$\frac{h^2\omega^2}{12}$
Jawna	0	0	0	$\frac{h^2\omega^2}{4}$	—
Różnice centralne	$\frac{1}{2}$	0	2	0	$-\frac{h^2\omega^2}{24}$
Foxa–Goodwina	$\frac{1}{2}$	$\frac{1}{12}$	2,45	0	$\mathcal{O}(h^3)$

W takiej sytuacji x oznacza wektor przemieszczeń węzłowych w bieżącej chwili czasowej, zaś m, b i k oznaczają macierze bezwładności, tłumień i sztywności.

Właściwości metody silnie zależą od wyboru parametrów α i β. Dla $\beta = \frac{1}{2}$ metoda

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

jest rzędu drugiego, więc błąd globalny jest proporcjonalny do kwadratu kroku, zaś stała błędu zależy silnie od wyboru parametru α. Dokładniej te informacje zostały przedstawione w tabeli 6.9. Oprócz informacji o stabilności tabela 6.9 zawiera informacje o wielkości błędów przy rozwiązywaniu równań opisujących drgania układów mechanicznych. Zawarto w niej błąd względny okresu oraz błąd bezwzględny amplitudy.

6.3. Szacowanie lokalnego błędu rozwiązania

Numeryczne rozwiązywanie równań różniczkowych zawsze jest obarczone niedokładnością. Zawsze pojawią się błędy, których wielkość jest zdeterminowana rzędem metody. W związku z tym należy odpowiedzieć na pytanie, czy da się oszacować wielkość lokalnego błędu rozwiązania w trakcie obliczeń. Od takiego oszacowania można uzależnić lokalną długość kroku całkowania równania, którą można zmieniać tak, by osiągać stałą wielkość błędu w całym przedziale rozwiązania.

Ponieważ najczęściej nie posiada się wiedzy *a priori* na temat dokładnego rozwiązanie analizowanego równania, konieczne staje się uzyskanie wartości odniesienia, która zostanie uznana za dokładną, w inny sposób. W obecnie stosowanych algorytmach jest używanych kilka metod. Poniżej zostanie naszkicowana chyba najłatwiejsza w szybkiej implementacji.

W zarysie algorytm można ująć następująco:

— wyznaczyć następną wartość rozwiązania równania $\tilde{y}_{i+1}$ z krokiem h,
— wyznaczyć wartość rozwiązania równania $\hat{y}_{i+1}$ z krokiem $\frac{1}{2}h$,
— na podstawie różnicy i wiedzy o rzędzie metody, można oszacować błąd.

Wartość $\hat{y}_{i+1}$ uzyskuje się, realizując dwa kroki rozwiązania wybraną metodą rozwiązywania równań różniczkowych. Jeżeli do rozwiązania zastosowano metodę rzędu p, błąd lokalny jest rzędu $p+1$. Zakłada się, że stała błędu nie zmienia się w przedziale $\langle x_i; x_{i+1}\rangle$ oraz błąd popełniony w każdym z kroków o długości $\frac{1}{2}h$ jest taki sam. Dla odpowiednio małej długości kroku to założenie jest spełnione. Przybliżone wartości rozwiązania w następnym kroku oblicza się na podstawie dokładnej wartości rozwiązania y_i. Zatem dokładna wartość rozwiązania może być zapisana jako

$$y_{i+1} \approx \tilde{y}_{i+1} + kh^{p+1} \tag{6.31}$$

$$y_{i+1} \approx \hat{y}_{i+1} + 2k\left(\frac{1}{2}h\right)^{p+1} \tag{6.32}$$

W drugim równaniu w wyrażeniu określającym błąd znajduje się czynnik 2, który wynika z faktu, że trzeba wykonać dwa kroki rozwiązania. Przyjmuje się, że w obu tych krokach błąd jest taki sam. Odejmując stronami równania (6.31) oraz (6.32), można wyznaczyć przybliżoną wartość błędu, która wynosi

$$kh^{p+1} \approx \frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{1 - \frac{1}{2^p}} = 2^p \frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{2^p - 1}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

co pozwala na uzyskanie następujących wyrażeń ujmujących to oszacowanie błędu

$$y_{i+1} \approx \tilde{y}_{i+1} + 2^p \frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{2^p - 1} \tag{6.33}$$

$$y_{i+1} \approx \hat{y}_{i+1} + \frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{2^p - 1} \tag{6.34}$$

Porównując zależności (6.31) oraz (6.32) z zależnościami (6.33) oraz (6.34), widać, że w zależności od tego, która długość kroku jest używana, uzyskuje się nieco inne wyrażenia na wielkość błędu. Opisane postępowanie związane z wyznaczaniem błędu znane jest jako zasada Rungego, jednak jest to typowe postępowanie ekstrapolacyjne.

Ponieważ rozwiązanie z mniejszą długością kroku jest dokładniejsze, jest ono przyjmowane jako wynik obliczeń. Natomiast lokalny błąd względny rozwiązania można oszacować za pomocą zależności

$$\hat{\varepsilon} = \left| \frac{\frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{2^p - 1}}{\hat{y}_{i+1}} \right| = \left| \frac{\hat{y}_{i+1} - \tilde{y}_{i+1}}{\hat{y}_{i+1}(2^p - 1)} \right|. \tag{6.35}$$

Gdy tak wyznaczony błąd przekracza wartość dopuszczalną, należy zmniejszyć krok na przykład dwa razy [29]. Gdy zaś błąd względny rozwiązania z krokiem $\frac{1}{2}h$ spada poniżej wartości dopuszczalnej, krok jest zwiększany dwa razy.

Porównując opisaną metodę z metodą Rungego–Kutty na przykład rzędu IV bez szacowania błędu lokalnego, można się przekonać, że wymaga ona wyznaczenia 11 wartości funkcji, zamiast 8, które należałoby wykonać przy kroku $\frac{1}{2}h$. Powoduje to, że należy wykonać około 1,375 raza więcej obliczeń. W zmian uzyskuje się możliwość oszacowania błędu lokalnego rozwiązania.

Wydajniejszą metodą szacowania błędu jest wykorzystanie tak par metod Rungego–Kutty. Opracowane między innymi przez: Fehlberga, Prince'a, Dormanda i innych metody nazywane osadzonymi pozwalają na szacowanie lokalnego błędu rozwiązania znacznie efektywniej. Metody te zostały opisane przy omawianiu metod Rungego–Kutty.

6.4. Metody rozwiązywania problemów brzegowych

Metody numeryczne stosowane do rozwiązywania problemów brzegowych, w zależności od sformułowania, można podzielić na dwie grupy – oparte na sformułowaniu: lokalnym i globalnym. Przedstawicielką pierwszej z grup jest metoda różnic skończonych, natomiast drugą reprezentuje metoda elementów skończonych. Pomijając różnice uwypuklone dalej w rozważaniach, można powiedzieć, że obie te metody polegają na:

— dyskretyzacji obszaru rozwiązania równania różniczkowego,
— przekształceniu równań różniczkowych do układu równań algebraicznych,
— rozwiązaniu układu równań algebraicznych.

Bardziej szczegółowo zostaną one omówione w rozdziałach 9 i 10.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.5. Wykorzystanie funkcji MATLAB-a

Aby rozwiązać układ równań różniczkowych rzędu pierwszego metodami Rungego–Kutty, które są zaimplementowane w MATLAB-ie jako `ode23` lub `ode45`, należy:

1. zapisać układ równań w postaci układu równań pierwszego rzędu,
2. zapisać układ równań w postaci m-pliku funkcyjnego,
3. wywołać funkcję `ode23` lub `ode45`.

M-plik funkcyjny opisujący układ równań różniczkowych musi przyjmować dwa argumenty (t, x), t – czas, zmienna niezależna, zaś x jest zmienną zależną, wyznaczaną w trakcie rozwiązywania równania. Wartością funkcji musi być kolumnowy wektor pochodnych $[z_1,\, z_2]^T$.

6.5.1. Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać układ równań różniczkowych rzędu pierwszego opisujących ruch masy zawieszonej sprężyście, przedstawionych w przykładzie z podrozdziału 6.2.4.

Rozwiązanie. Układ równań różniczkowych pierwszego rzędu, który należy rozwiązać, wygląda następująco

$$\begin{bmatrix} \dot{z_1} \\ \dot{z_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{8}{5} & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Powyższy układ równań należy zapisać w postaci funkcji, którą będzie wywoływała wykorzystana metoda rozwiązywania równań różniczkowych. Najwygodniej będzie użyć m-plików funkcyjnych. M-plik funkcyjny o nazwie `uklad1.m` może mieć następującą postać:

```
  function dy=uklad1(t,z)
2 dy=[0 1;-8/5 -1/5]*z;
```

Interpretacja wyników zależy od przyjętych podstawień. W rozwiązywanym przykładzie z_1 oznacza przemieszczenie masy, oznaczane w równaniu ruchu (6.14) przez x. Zaś z_2 oznacza jej prędkość, którą w równaniu oznaczono przez $\dot{x}$. Rozwiązanie równania zapisanego w postaci m-pliku funkcyjnego w przedziale czasu $t \in \langle 0; 6\rangle$ wykonuje się za pomocą polecenia

`[T,Z]=ode45(@uklad1,[0,6],[0.1,0]);`

Wynikowa macierz `T` zawiera wartości zmiennej niezależnej, którą w tym wypadku jest czas. W tych chwilach czasowych wyznaczane były wartości rozwiązania. W pierwszej kolumnie zwracanej macierzy `Z` zapisane są wartości z_1, w drugiej kolumnie z_2. Każdej chwili czasowej odpowiada jeden wiersz macierzy `Z`. Następnie można narysować rozwiązania, posługując się poleceniem `plot(T,Z(:,1))` dla przemieszczeń lub `plot(T,Z(:,2))` dla prędkości.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

6.6. Podsumowanie

Podsumowanie rozdziału o rozwiązywaniu równań różniczkowych powinno zawierać wskazówki dotyczące wyboru metod. Niestety, pomimo wielkiej liczby publikacji dotyczących tych zagadnień, nie da się podać ogólnej recepty, która dotyczyłaby wszystkich spotykanych problemów. Można wskazać metody nadające się szczególnie do pewnych zagadnień. Można zwrócić uwagę na pewne zalety czy wady poszczególnych metod, które mogłyby uzasadniać ich użycie.

W ostatnim czasie bardzo popularne stały się metody Rungego–Kutty. Nie bez powodu są chyba najczęściej wykorzystywane przy obliczeniach numerycznych do modelowana mechanicznych układów dyskretnych. Producent MATLAB-a sugeruje metodę Rungego–Kutty rzędów IV i V jako podstawową metodę przy rozwiązywaniu równań różniczkowych. Dopiero gdy ona zawiedzie, zaleca się poszukiwanie innej metody. Rosnący z rzędem obszar stabilności absolutnej, wydajne, zoptymalizowane procedury, łatwa kontrola błędu lokalnego to tylko niektóre z cech odpowiadających za dużą popularność tych metod.

Innym powodem ich popularności jest to, że nie są tak wrażliwe na nieciągłości czy szybkie zmiany funkcji wymuszających jak metody Adamsa. Przy częstym znacznym zapasie mocy obliczeniowej dzisiejszych komputerów mniej istotna staje się wydajność metod Adamsa, a cenniejsza jest uniwersalność metod Rungego–Kutty. Uwzględniając przydatność niejawnych metod Rungego–Kutty do rozwiązywania równań sztywnych, można powiedzieć, że są to w tej chwili najskuteczniejsze narzędzia do numerycznego rozwiązywania równań różniczkowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 7

Rozwiązywanie układów liniowych równań algebraicznych

7.1. Wstęp

Numeryczna algebra liniowa, zajmująca się układami równań liniowych, jest niezwykle ważnym działem analizy numerycznej. Wynika to między innymi z jej fundamentalnej roli w rozwiązywaniu współczesnych problemów inżynierskich. Nie jest zatem zaskoczeniem, że wiele metod numerycznych służy do rozwiązywania układów algebraicznych równań liniowych. Jest chyba najlepiej rozwiniętym działem analizy numerycznej i dalej przeżywa dynamiczny rozwój.

Rozwiązywanie problemów inżynierskich i naukowych metodami elementów skończonych, różnic skończonych, elementów brzegowych, objętości skończonych i wielu innych bezpośrednio prowadzi do układów liniowych równań algebraicznych. Zatem rozwiązywanie większości współczesnych problemów inżynierskich sprowadza się do wyznaczenia rozwiązania takich układów równań.

7.1.1. Układy równań liniowych

W niniejszym rozdziale przyjmuje się, iż należy rozwiązać następujący układ równań liniowych, w którym wszystkie współczynniki a_{ij}, wartości b_i i rozwiązania x_i są rzeczywiste

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n &= b_3 \\ \vdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n &= b_n \end{cases} \tag{7.1}$$

Taki układ równań da się zapisać w postaci macierzowej

$$\mathbb{A}\mathbb{x} = \mathbb{b} \tag{7.2}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie poszczególne macierze to

$$\mathbb{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}, \quad \mathbb{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{zaś } \mathbb{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

Zakłada się, że macierz $\mathbb{A}$ jest kwadratowa i jej rząd jest równy n oraz, że rozwiązanie układu istnieje i jest oznaczane przez $\mathbb{x}^*$. Taki liniowy układ równań algebraicznych można rozwiązywać metodami należącymi do jednej z dwóch głównych grup:

— metod bezpośrednich lub
— metod iteracyjnych.

Pomijając fakt istnienia dużej liczby odmian każdej z metod, można powiedzieć, że metody bezpośrednie stosuje się do małych układów – zawierających niewielką liczbę niewiadomych i ewentualnie wiele wektorów obciążeń. Zasadniczą korzyścią jest to, że w takim przypadku, za pomocą eliminacji Gaussa, można jednocześnie rozwiązywać układ z wieloma prawymi stronami [50]. Daje to wielką korzyść przy badaniu zachowania układu przy np. zmianie wymuszeń.

Przy rozwiązywaniu układów równań liniowych metodami iteracyjnymi niezbędny jest sposób określania wielkości błędu albo, inaczej mówiąc, niedokładności rozwiązania układu równań. Jest to o tyle trudne, że każde przybliżenie rozwiązania to macierz kolumnowa. Zatem nie można wykorzystać zwykłej różnicy między dwoma kolejnymi przybliżeniami rozwiązania. Należy najpierw określić sposób na przyporządkowanie wektorowi (macierzy kolumnowej) liczby jednoznacznie go charakteryzującej.

7.1.2. Normy wektorów i macierzy

Normy pozwalają na jednoznaczne przyporządkowanie liczby wektorowi. Najprostsza w interpretacji jest norma euklidesowa, która oznacza po prostu długość wektora w przestrzeni n-wymiarowej. Stawiane normom warunki można w skrócie ująć następująco: norma wektora powinna wynosić zero dla wektora zerowego, mnożenie wektora przez skalar ma sprowadzać się do mnożenia normy wektora przez skalar oraz prawdziwa ma być nierówność trójkąta, którą można zapisać

$$||\mathbb{x} + \mathbb{y}|| \leqslant ||\mathbb{x}|| + ||\mathbb{y}||$$

Najczęściej stosowane normy wymieniono poniżej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Norma euklidesowa, zwana też normą drugą, określana jest jako pierwiastek z sumy kwadratów składowych wektora

$$||\mathbb{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Norma-1 to suma wartości bezwzględnych składowych wektora

$$||\mathbb{x}||_1 = \sum_{i=1}^{n} |x_i|$$

Norma-∞ zwana też normą-maksimum to maksymalna co do modułu składowa wektora

$$||\mathbb{x}||_\infty = \max_{i=1,\ldots,n} |x_i|$$

Uogólniając pojęcie normy wektora, wprowadza się pojęcie indukowanej normy macierzy. Normy wektorów indukują odpowiednie normy macierzy [22, 23, 48]. Indukowaną normę macierzy można zdefiniować jako

$$||\mathbb{A}||_p = \sup_{||\mathbb{x}||_p=1} ||\mathbb{A}\mathbb{x}||_p \tag{7.3}$$

gdzie $p \in \{1, 2, \infty\}$, zaś sup oznacza kres górny.

Najważniejsze indukowane normy macierzy wymieniono poniżej.

Norma euklidesowa wektorów indukuje normę spektralną macierzy

$$||\mathbb{A}||_2 = \sqrt{\lambda_{max}}$$

której wartość jest równa pierwiastkowi kwadratowemu z największej wartości własnej macierzy $\mathbb{A}$.

Norma-1 macierzy jest określona następująco

$$||\mathbb{A}||_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$$

Norma-∞ wyrażona wzorem

$$||\mathbb{A}||_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$$

Norma Frobeniusa macierzy $\mathbb{A}_{m\times n}$ jest opisywana zależnością

$$||\mathbb{A}||_F = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij}^2}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Norma Frobeniusa nie jest indukowana przez żadną normę wektorową, ale jest bardzo wygodna w obliczaniu.

Normy macierzy i wektorów pozwalają określić warunki skończenia obliczeń w wygodny i zwięzły sposób.

Warunki skończenia obliczeń. Przy metodach bezpośrednich nie używa się takich warunków, natomiast przy metodach iteracyjnych są one konieczne. Zazwyczaj są to warunki typu

$$\frac{||\mathbb{A}\mathbb{x}_i - \mathbb{b}||}{||\mathbb{b}||} \leqslant \varepsilon \tag{7.4}$$

gdzie ε oznacza dopuszczalny błąd. Licznik powyższego ułamka oznacza residuum (albo inaczej resztę) rozwiązania, przy rozwiązaniu dokładnym powinno ono wynosić zero. Zazwyczaj nie jest istotne, jaka norma zostanie użyta do wyznaczania wartości wyrażenia (7.4). Ale od wyboru normy może zależeć dopuszczalna wielkość błędu ε.

Uwarunkowanie układu równań pozwala ocenić, jak wrażliwy na błędy zaokrągleń będzie dany układ równań. Różne problemy numeryczne mogą być źle lub dobrze uwarunkowane. Jak już podano w podrozdziale 1.3, złe uwarunkowanie problemu nastręcza problemów z rozwiązaniem niezależnie od wybranej metody rozwiązania i powoduje, że niewielkie błędy zaokrągleń mają istotny wpływ na wynik końcowy.

Dla niektórych zagadnień numerycznych określono wskaźnik uwarunkowania, który pozwala ocenić, czy dany problem będzie wrażliwy na błędy, czy też nie. Dla rozwiązywania układów równań takim wskaźnikiem uwarunkowania jest zależność

$$\varkappa(\mathbb{A}) = ||\mathbb{A}|| \cdot ||\mathbb{A}^{-1}||$$

Wartość tego wskaźnika zależy od użytej do obliczenia go normy, ale zawsze jest co najmniej równa jedności. Zatem układy równań, których macierze współczynników mają wskaźnik uwarunkowania najbliższy jedności, są najmniej wrażliwe na błędy zaokrągleń, co widać z zależności (7.5).

Błędy zaokrągleń i jakość rozwiązania układu równań $\mathbb{A}\mathbb{x} = \mathbb{b}$ można analizować, biorąc pod uwagę odwrotność macierzy $\mathbb{A}^{-1}$ zawierającą wartości obarczone błędem czy zaburzoną prawą stronę układu równań $\mathbb{b}$. W sytuacji gdy zaburzona jest odwrotność $\mathbb{A}^{-1?} \approx \mathbb{A}^{-1}$, można zapisać równania

$$\mathbb{x} = \mathbb{A}^{-1}\mathbb{b} \quad \text{oraz} \quad \widehat{\mathbb{x}} = \mathbb{A}^{-1?}\mathbb{b}$$

oraz wyznaczyć wielkość błędu rozwiązania układu równań w takiej sytuacji

$$||\mathbb{x} - \widehat{\mathbb{x}}|| = ||\mathbb{x} - \mathbb{A}^{-1?}\mathbb{b}|| = ||\mathbb{x} - \mathbb{A}^{-1?}\mathbb{A}\mathbb{x}|| = ||(\mathbb{I} - \mathbb{A}^{-1?}\mathbb{A})\mathbb{x}||$$

Na mocy nierówności trójkąta można powyższą równość zapisać w postaci

$$||\mathbb{x} - \widehat{\mathbb{x}}|| \leqslant ||\mathbb{I} - \mathbb{A}^{-1?}\mathbb{A}|| \cdot ||\mathbb{x}||$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Dzieląc obustronnie przez $||\mathbb{x}||$, uzyskuje się oszacowanie zaburzenia względnego rozwiązania wyrażone przez wielkość zaburzenia odwrotności macierzy $\mathbb{A}$.

$$\frac{||\mathbb{x} - \widehat{\mathbb{x}}||}{||\mathbb{x}||} \leqslant ||\mathbb{I} - \mathbb{A}^{-1?}\mathbb{A}||$$

Chociaż przy rozwiązywaniu układów równań liniowych nie dokonuje się jawnego odwracania macierzy, to powyższy wynik pokazuje, że błędy w operacjach na wierszach macierzy mają duży wpływ na rozwiązanie.

Z kolei, gdy zaburzona jest prawa strona układu równań, i zamiast $\mathbb{b}$ występuje wektor $\widehat{\mathbb{b}}$, błąd rozwiązania można oszacować w sposób następujący: oznaczając zaburzone rozwiązanie przez $\widehat{\mathbb{x}}$, można napisać równania

$$\mathbb{A}\mathbb{x} = \mathbb{b} \quad \text{oraz} \quad \mathbb{A}\widehat{\mathbb{x}} = \widehat{\mathbb{b}}$$

Wobec tego można wyznaczyć wielkość błędu bezwzględnego zgodnie z następującym ciągiem logicznym wykorzystującym nierówność trójkąta

$$||\mathbb{x} - \widehat{\mathbb{x}}|| = ||\mathbb{A}^{-1}\mathbb{b} - \mathbb{A}^{-1}\widehat{\mathbb{b}}|| = ||\mathbb{A}^{-1}(\mathbb{b} - \widehat{\mathbb{b}})|| \leqslant ||\mathbb{A}^{-1}|| \cdot ||\mathbb{b} - \widehat{\mathbb{b}}||$$

Aby wyznaczyć wielkość błędu względnego, można wykorzystać fakt, że $\mathbb{A}\mathbb{x} = \mathbb{b}$. Zatem prawą stronę nierówności można pomnożyć przez $||\mathbb{A}\mathbb{x}||$ i podzielić przez $||\mathbb{b}||$, nie powodując żadnych zaburzeń. W wyniku otrzymuje się zależność

$$||\mathbb{x} - \widehat{\mathbb{x}}|| \leqslant ||\mathbb{A}^{-1}|| \cdot ||\mathbb{b} - \widehat{\mathbb{b}}|| = ||\mathbb{A}^{-1}|| \cdot ||\mathbb{A}\mathbb{x}|| \frac{||\mathbb{b} - \widehat{\mathbb{b}}||}{||\mathbb{b}||}$$

co można zapisać jako

$$||\mathbb{x} - \widehat{\mathbb{x}}|| \leqslant ||\mathbb{A}^{-1}|| \cdot ||\mathbb{A}|| \cdot ||\mathbb{x}|| \frac{||\mathbb{b} - \widehat{\mathbb{b}}||}{||\mathbb{b}||}$$

Wobec tego, wykorzystując wskaźnik uwarunkowania, można zapisać zależność na wielkość błędu względnego jako

$$\frac{||\mathbb{x} - \widehat{\mathbb{x}}||}{||\mathbb{x}||} \leqslant ||\mathbb{A}^{-1}|| \cdot ||\mathbb{A}|| \frac{||\mathbb{b} - \widehat{\mathbb{b}}||}{||\mathbb{b}||} = \varkappa(\mathbb{A}) \frac{||\mathbb{b} - \widehat{\mathbb{b}}||}{||\mathbb{b}||}$$

gdzie oznaczenie $\varkappa(\mathbb{A})$ oznacza wskaźnik uwarunkowania macierzy $\mathbb{A}$.

Uogólniając powyższe rozumowanie i wprowadzając oznaczenia wektora residualnego $r = \mathbb{b} - \mathbb{A}\widehat{\mathbb{x}}$ oraz wektora błędu $e = \mathbb{x} - \widehat{\mathbb{x}}$, powiązanych zależnością

$$\mathbb{A}e = r$$

można napisać następujący wzór [23]

$$\frac{1}{\varkappa(\mathbb{A})} \frac{||r||}{||b||} \leqslant \frac{||e||}{||\mathbb{x}||} \leqslant \varkappa(\mathbb{A}) \frac{||r||}{||b||} \tag{7.5}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

pozwalający oszacować wielkość błędu popełnianego przy rozwiązaniu układu algebraicznych równań liniowych.

Wartości i wektory własne macierzy $\mathbb{A}$ stanowią rozwiązanie problemu zapisywanego w następujący sposób

$$(\mathbb{A} - \lambda\mathbb{I})\mathbb{x} = 0$$

gdzie λ to wartość własna, zaś $\mathbb{x}$ jest wektorem własnym. Każdej wartości własnej odpowiada nieskończenie wiele wektorów własnych. W mechanice najczęściej rozwiązuje się tego typu zadania przy wyznaczaniu częstości drgań własnych i postaci drgań własnych układu. Wtedy, zależnie od sformułowania, wartość własna $\lambda = \omega^2$ jest kwadratem częstości drgań własnych lub jego odwrotnością $\lambda = \frac{1}{\omega^2}$. Natomiast wektor własny opisuje postać drgań układu.

7.2. Metody bezpośrednie

Rodzina metod bezpośrednich, polegających na przekształceniu macierzy współczynników $\mathbb{A}$ do postaci diagonalnej lub trójkątnej przez sumowanie poszczególnych wierszy mnożonych przez czynnik od nazwiska twórcy, nazywa się metodami Gaussa.

7.2.1. Eliminacja Gaussa

Jedną z podstawowych metod rozwiązywania układów równań liniowych jest metoda eliminacji Gaussa, która, najprościej mówiąc, polega na przekształceniu macierzy $\mathbb{A}$ do macierzy trójkątnej górnej. Jest chyba najogólniejszą metodą rozwiązywania układów równań liniowych. Niestety, wymaga znacznego nakładu obliczeniowego. W związku z tym rozwinięto wiele pochodnych metod, które pozwalają zmniejszyć liczbę obliczeń konieczną do uzyskania rozwiązania. Oczywiście, nic nie dzieje się bez konsekwencji. Mniejsza liczba operacji przy rozwiązywaniu powoduje, że macierz współczynników musi mieć pewne szczególne cechy.

Metoda eliminacji Gaussa polega na przekształceniu macierzy $\mathbb{A}$ do postaci trójkątnej górnej. Ten zabieg wykonuje się, mnożąc poszczególne wiersze przez wybraną wartość i następnie sumując z innymi wierszami. Zabieg ten ma doprowadzić do wyzerowania elementów pod główną przekątną macierzy $\mathbb{A}$. Opisane elementarne operacje na macierzy można wykonać, mnożąc macierz $\mathbb{A}$ przez odpowiednią macierz przekształcenia elementarnego. Aby zilustrować to stwierdzenie, zostanie przedstawionych kilka przykładów takich działań elementarnych wraz z odpowiednimi macierzami przekształcenia.

Mnożenie wiersza przez skalar można zapisać w postaci przekształcenia

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \lambda a_{21} & \lambda a_{22} & \lambda a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zamianę kolejności wierszy można uzyskać przez mnożenie następujących macierzy

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{31} & a_{32} & a_{33} \\ a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \end{bmatrix}$$

Mnożenie wiersza przez wartość i dodanie do innego uzyskać można używając następującego mnożenia danej macierzy przez macierz przekształcenia

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \lambda & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \lambda a_{11} + a_{31} & \lambda a_{12} + a_{32} & \lambda a_{13} + a_{33} \end{bmatrix}$$

W podobny sposób można uzyskać odpowiednie działania na wektorach, zaś macierze przekształceń elementarnych dla wektorów mają tę samą postać co dla macierzy kwadratowych.

Przekształcenie macierzy do macierzy jednostkowej czy macierzy trójkątnej dolnej lub trójkątnej górnej wymaga wielu przekształceń elementarnych wykonywanych jedno po drugim. Taki ciąg przekształceń można zapisać jako iloczyn odpowiednich macierzy przekształceń.

W ten sposób można też wyznaczyć macierz odwrotną, jeżeli zna się ciąg przekształceń prowadzących do przekształcenia macierzy $\mathbb{A}$ w macierz jednostkową [23]. Po wyjaśnieniu związku elementarnych przekształceń macierzy z macierzami przekształceń można przejść do omawiania eliminacji Gaussa.

Eliminacja Gaussa zostanie przedstawiona na przykładzie poniższego układu algebraicznych równań liniowych

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 3 & -1 & 3 & 2 \\ 4 & -4 & 4 & -2 \\ 3 & -5 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 11 \\ 10 \\ -18 \\ 0 \end{bmatrix}$$

W tabeli 7.1 w pierwszej kolumnie przedstawiono macierz współczynników przed eliminacją kolumny macierzy, w drugiej opis wykonywanych działań, przy czym w_i oznacza i-ty wiersz macierzy, kolumna trzecia przedstawia macierz $\mathbb{A}$ po wykonaniu opisanych operacji. Ostatnia kolumna przedstawia wektor prawych stron po wykonaniu danego kroku eliminacji.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 7.1. Kolejne kroki eliminacji Gaussa

$\mathbb{A}$	Działanie	$\mathbb{A}_{mod}$	$\mathbb{b}_{mod}$
$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 3 & -1 & 3 & 2 \\ 4 & -4 & 4 & -2 \\ 3 & -5 & 1 & 2 \end{bmatrix}$	$w_2 = w_2 - 3 \cdot w_1$, $w_3 = w_3 - 4 \cdot w_1$, $w_4 = w_4 - 3 \cdot w_1$	$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & -8 & 8 & -6 \\ 0 & -8 & 4 & -1 \end{bmatrix}$	$\begin{bmatrix} 11 \\ -23 \\ -62 \\ -33 \end{bmatrix}$
$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & -8 & 8 & -6 \\ 0 & -8 & 4 & -1 \end{bmatrix}$	$w_3 = w_3 - 2w_2$, $w_4 = w_4 - 2w_2$	$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & -8 & 1 \end{bmatrix}$	$\begin{bmatrix} 11 \\ -23 \\ -16 \\ 13 \end{bmatrix}$
$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & -8 & 1 \end{bmatrix}$	$w_4 = w_4 - 2w_3$	$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & 9 \end{bmatrix}$	$\begin{bmatrix} 11 \\ -23 \\ -16 \\ 45 \end{bmatrix}$

Współczynniki znajdujące się w drugiej kolumnie tabeli 7.1, opisujące przeprowadzone działania mogą zostać wykorzystane do wykonania rozkładu LU (patrz podrozdział 7.3.1). Zatem po przeprowadzeniu eliminacji Gaussa układ równań przyjmuje postać

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 11 \\ -23 \\ -16 \\ 45 \end{bmatrix} \tag{7.6}$$

Następnym etapem rozwiązywania układu równań powinno być podstawianie wstecz. Polega ono na rozwiązaniu uzyskanego układu niejako „od tyłu", czyli zaczynając od ostatniego równania, które w tym przypadku wygląda następująco

$$9x_4 = 45$$

co daje wartość $x_4 = \frac{45}{9} = 5$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Równanie zapisane w wierszu przedostatnim ma postać

$$-4x_3 - 4x_4 = -16$$

czyli $-4x_3 - 4 \cdot 5 = -16$, co daje wynik $x_3 = \frac{-16+4\cdot 5}{-4} = -1$

Podobnie rozwiązuje się pozostałe równania przekształconego układu równań. Za każdym razem do wyznaczenia jest tylko jedna niewiadoma, co pozwala szybko uzyskać pełne rozwiązanie układu. Zatem końcowa postać rozwiązanie jest następująca

$$\mathbb{x} = \begin{bmatrix} 2 & 3 & -1 & 5 \end{bmatrix}^T$$

Gdy macierz współczynników $\mathbb{A}$ jest macierzą trójkątną dolną, można opracować podobny algorytm rozwiązania układu równań, zaczynając od pierwszego równania układu równań, które wprost daje wartość pierwszej niewiadomej. Następnie przechodząc kolejne równania układu, od góry do dołu, w każdym kroku rozwiązuje się równanie z jedną niewiadomą. Ten tok postępowania nazywany jest podstawianiem w przód.

Zatem można powiedzieć, że układ równań z macierzą współczynników trójkątną dolną lub górną jest łatwy do rozwiązania. Podobnie jak układ, w którym macierz $\mathbb{A}$ jest macierzą diagonalną.

Przedstawiony powyżej wariant eliminacji Gaussa znajduje zastosowanie w obliczeniach analitycznych. Przy numerycznym rozwiązywaniu układów równań liniowych należy przeprowadzać częściowy lub pełny wybór elementu głównego. Częściowy wybór elementu głównego polega na tym, że zamienia się kolejność wierszy tak, by na głównej przekątnej znajdował się maksymalny co do wartości bezwzględnej element kolumny.

7.2.2. Metoda Gaussa–Jordana

Metoda Gaussa–Jordana polega na tym, że przeprowadza się eliminację nie tylko elementów leżących pod główną przekątną, lecz z całej kolumny oprócz głównej przekątnej. W efekcie uzyskuje się macierz diagonalną, czyli macierz z elementami niezerowymi na głównej przekątnej.

7.3. Metody polegające na rozkładach macierzy

Chociaż można je postrzegać jako całkowicie oddzielną grupę metod, w rzeczywistości są pewnymi odmianami eliminacji Gaussa.

7.3.1. Rozkład LU

Jak już zauważono wcześniej, bardzo wygodną, ze względu na łatwość rozwiązywania układu równań liniowych (7.2), postacią macierzy współczynników $\mathbb{A}$ jest postać trójkątna górna lub trójkątna dolna. Macierz współczynników można też wyrazić w postaci iloczynu macierzy trójkątnej dolnej i górnej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Niestety nie każdą macierz można wyrazić w postaci iloczynu takich dwóch macierzy:

— trójkątnej dolnej, oznaczanej jako $\mathbb{L}$ oraz
— trójkątnej górnej oznaczanej przez $\mathbb{U}$.

Jedna z metod realizacji rozkładu LU wynika wprost z metody eliminacji Gaussa. Macierz $\mathbb{A}$ po przeprowadzeniu eliminacji przyjmuje postać macierzy trójkątnej górnej $\mathbb{U}$, co zostało przedstawione w równaniu (7.6). Współczynniki, których używano przy eliminacji, a które zamieszczono w drugiej kolumnie tabeli 7.1 na stronie 158, należy zapisać w odpowiednich kolumnach macierzy. W ten sposób powstaje macierz trójkątna dolna $\mathbb{L}$, która wygląda następująco

$$\mathbb{U} = \begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -4 & 6 & -1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & 9 \end{bmatrix}, \qquad \mathbb{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 4 & 2 & 1 & 0 \\ 3 & 2 & 2 & 1 \end{bmatrix}$$

Jedynki na głównej przekątnej macierzy $\mathbb{L}$ oznaczają, że te elementy nie były modyfikowane. W ten sposób przy okazji eliminacji Gaussa zrealizowano rozkład macierzy $\mathbb{A}$ na macierz trójkątną górną $\mathbb{U}$ oraz macierz trójkątną dolną $\mathbb{L}$. Jedynki na głównej przekątnej w macierzy $\mathbb{L}$ oznaczają, że uzyskano rozkład Doolittle'a, w odróżnieniu od rozkładu Crouta, który jedynki ma na przekątnej macierzy $\mathbb{U}$. Inny ważny rozkład, rozkład Choleskiego daje $\mathbb{U} = \mathbb{L}^T$, ale wymaga macierzy symetrycznych, dodatnio określonych.

Jeżeli jednak daną macierz da się rozłożyć na macierze $\mathbb{L}$ i $\mathbb{U}$, można przyspieszyć proces rozwiązywania układu równań. A zatem, rozwiązując układ równań w postaci

$$\mathbb{A}\mathbb{x} = \mathbb{b}$$

można rozłożyć macierz $\mathbb{A} = \mathbb{L}\mathbb{U}$. Dzięki rozłożeniu macierzy współczynników na takie dwie macierze uzyskuje się następującą postać układu równań

$$\mathbb{L}\mathbb{U}\mathbb{x} = \mathbb{b}$$

Po podstawieniu $\mathbb{U}\mathbb{x} = \mathbb{z}$ uzyskuje się układ równań

$$\mathbb{L}\mathbb{z} = \mathbb{b}$$

który bardzo szybko można rozwiązać metodą podstawiania w przód. Następnie, skoro znane są wartości $\mathbb{z}$, rozwiązuje się układ wynikający z podstawienia

$$\mathbb{U}\mathbb{x} = \mathbb{z}$$

Ponieważ macierz $\mathbb{U}$ jest macierzą trójkątną górną, ten układ również bardzo szybko można rozwiązać metodą podstawiania wstecz.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

7.3.2. Rozkład QR

Rozkład LU pozwalał szybko rozwiązać układ równań liniowych. Podobnie rozkład QR umożliwia bardzo szybkie rozwiązywanie układu równań w postaci

$$\mathbb{A}\mathbb{x} = \mathbb{b}$$

Macierz współczynników $\mathbb{A}$ jest rozkładana na iloczyn dwóch macierzy

$$\mathbb{A} = \mathbb{Q}\mathbb{R},$$

gdzie $\mathbb{Q}$ jest macierzą ortogonalną, natomiast $\mathbb{R}$ jest macierzą trójkątną górną. Macierz ortogonalna to macierz spełniająca warunek $\mathbb{Q}\mathbb{Q}^T = \mathbb{I}$. Co oznacza, że transpozycja macierzy ortogonalnej jest tożsama z jej odwróceniem, czyli $\mathbb{Q}^{-1} = \mathbb{Q}^T$. Rozwiązywanie układu równań przebiega analogicznie do rozwiązania z rozkładem LU. Dzięki rozłożeniu macierzy współczynników na dwie macierze $\mathbb{Q}$ i $\mathbb{R}$ uzyskuje się następującą postać układu równań

$$\mathbb{Q}\mathbb{R}\mathbb{x} = \mathbb{b}$$

Po podstawieniu $\mathbb{U}\mathbb{x} = \mathbb{z}$ uzyskuje się układ równań

$$\mathbb{Q}\mathbb{z} = \mathbb{b}$$

który bardzo szybko można rozwiązać. Dzięki temu, że $\mathbb{Q}$ jest macierzą ortogonalną, wystarczy tylko mnożenie macierzowe

$$\mathbb{Q}^{-1}\mathbb{Q}\mathbb{z} = \mathbb{Q}^{-1}\mathbb{b}$$

czyli można zapisać wyrażenie na $\mathbb{z}$

$$\mathbb{z} = \mathbb{Q}^T\mathbb{b}$$

Takie obliczenia można bardzo szybko wykonać. Następnie, skoro znane są wartości $\mathbb{z}$, rozwiązuje się układ wynikający z podstawienia

$$\mathbb{R}\mathbb{x} = \mathbb{z}$$

Ponieważ macierz $\mathbb{R}\mathbb{x}$ jest macierzą trójkątną górną, ten układ również bardzo szybko można rozwiązać, stosując metodę podstawiania wstecz.

Istnieje kilka metod realizacji rozkładu QR. Jednym ze sposobów uzyskania rozkładu QR jest ortogonalizacja Grama–Schmidta, która pozwala na znalezienie wektorów ortonormalnych dla zadanego układu wektorów liniowo niezależnych. Bardzo ważne jest to, że rozkładu QR można dokonać dla każdej macierzy odwracalnej, czyli metodę można stosować dla każdego układu równań, który ma jednoznaczne rozwiązanie. Natomiast nie każdą macierz nieosobliwą można rozłożyć na macierze $\mathbb{L}$ i $\mathbb{U}$. Inny sposób otrzymania takiego rozkładu polega na zastosowaniu przekształcenia Householdera. Taki algorytm jest najprostszym ze stabilnych algorytmów numerycznych stosowanych do rozkładu QR. Najbardziej złożonym i chyba najskuteczniejszym algorytmem realizacji rozkładu QR są obroty Givensa.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

7.4. Metody iteracyjne

7.4.1. Metoda Gaussa–Seidla

Metoda Gaussa–Seidla należy do metod iteracyjnych rozwiązywania układów równań liniowych, zaś jej idea jest bliska idei metody iteracji prostej znanej z poprzednich rozdziałów. Można ją stosować do macierzy przekątniowo dominujących, w których elementy na głównej przekątnej mają wartości większe niż pozostałe elementy wiersza. Dominacja elementów leżących na głównej przekątnej zapewnia zbieżność metody Gaussa–Seidla. Jej interpretacja jest bardzo intuicyjna, łatwo to zauważyć, gdy rozwiązywany układ ma dwa równania. Metoda sprowadza się do wyznaczenia każdej niewiadomej oddzielnie i następnie iteracyjnie poprawia się początkowe przybliżenie rozwiązania.

W metodzie Gaussa–Seidla macierz $\mathbb{A}$ przedstawia się w postaci sumy macierzy trójkątnej górnej $\mathbb{U}$ z zerami na głównej przekątnej, trójkątnej dolnej $\mathbb{L}$ z zerami na głównej przekątnej oraz macierzy diagonalnej $\mathbb{D}$ zawierającej główną przekątną. Taki rozkład sprowadza się do zapisania odpowiednich elementów macierzy $\mathbb{A}$ w odpowiednich macierzach $\mathbb{L}$, $\mathbb{D}$ i $\mathbb{U}$

$$\mathbb{A} = \mathbb{L} + \mathbb{D} + \mathbb{U} \tag{7.7}$$

Można zapisać zatem

$$(\mathbb{L} + \mathbb{D} + \mathbb{U})\mathbb{x} = \mathbb{b} \tag{7.8}$$

Po wymnożeniu oraz przeniesieniu na prawą stronę czynnika $\mathbb{U}\mathbb{x}$ uzyska się równanie w postaci

$$(\mathbb{L} + \mathbb{D})\mathbb{x} = -\mathbb{U}\mathbb{x} + \mathbb{b} \tag{7.9}$$

Następnym krokiem jest lewostronne pomnożenie równania przez czynnik $(\mathbb{L} + \mathbb{D})^{-1}$

$$\mathbb{x} = -(\mathbb{L} + \mathbb{D})^{-1}\mathbb{U}\mathbb{x} + (\mathbb{L} + \mathbb{D})^{-1}\mathbb{b} \tag{7.10}$$

Uzupełnienie uzyskanej zależności indeksami pomaga sprecyzować kolejność przeprowadzanych obliczeń

$$\mathbb{x}_{i+1} = -(\mathbb{L} + \mathbb{D})^{-1}\mathbb{U}\mathbb{x}_i + (\mathbb{L} + \mathbb{D})^{-1}\mathbb{b} \tag{7.11}$$

Obliczenia prowadzi się, rozpoczynając od pewnego wektora startowego $\mathbb{x}_0$ aż do uzyskania zadowalającego przybliżenia rozwiązania. Ponieważ macierze $(\mathbb{L} + \mathbb{D})^{-1}\mathbb{U}$ oraz $(\mathbb{L}+\mathbb{D})^{-1}$ wylicza się tylko raz, metoda pozwala na stosunkowo szybkie uzyskanie rozwiązania.

Przykład ilustrujący powyższe rozważania pozwoli lepiej zrozumieć ideę omawianej metody. Należy rozwiązać poniższy układ równań zapisany w postaci macierzowej.

$$\begin{bmatrix} 4 & 1 & -1 & -1 \\ 1 & -4 & 2 & -1 \\ 1 & -2 & 8 & -1 \\ 1 & -2 & -3 & 6 \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -9 \\ 1 \\ 11 \end{bmatrix} \tag{7.12}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Ponieważ macierz $\mathbb{A}$ jest dominująca przekątniowo (współczynniki o największej wartości bezwzględnej leżą na głównej przekątnej) nie ma potrzeby zamieniania kolejności równań. Rozkład macierzy współczynników $\mathbb{A}$ na sumę składników $\mathbb{L}$, $\mathbb{D}$ i $\mathbb{U}$ jest przedstawiony poniżej

$$\mathbb{L} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 \\ 1 & -2 & -3 & 0 \end{bmatrix}, \quad \mathbb{D} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}, \quad \mathbb{U} = \begin{bmatrix} 0 & 1 & -1 & -1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Poszczególne składniki zależności (7.11) to

$$\mathbb{Q} = (\mathbb{L} + \mathbb{D})^{-1}\mathbb{U} = \begin{bmatrix} 0 & 0{,}250000 & -0{,}25000 & -0{,}250000 \\ 0 & 0{,}062500 & -0{,}56250 & 0{,}187500 \\ 0 & -0{,}015625 & -0{,}10938 & -0{,}046875 \\ 0 & -0{,}028646 & -0{,}20052 & 0{,}080729 \end{bmatrix}$$

$$\mathbb{C} = (\mathbb{L} + \mathbb{D})^{-1}\mathbb{b} = \begin{bmatrix} -0{,}50000 \\ 2{,}12500 \\ 0{,}71875 \\ 2{,}98440 \end{bmatrix}$$

Zależność pozwalająca na rozwiązanie zadanego układu równań przyjmuje teraz następującą postać

$$\mathbb{x}_{i+1} = -\mathbb{Q}\mathbb{x}_i + \mathbb{C}$$

Przyjmując punkt startowy zerowy, czyli $\mathbb{x}_0 = [0\,0\,0\,0]^T$, wykonywane są kolejne obliczenia i dają one ciąg kolejnych przybliżeń (tab. 7.2). Oprócz wektorów stanowiących kolejne przybliżenia rozwiązania tabela 7.2 zawiera wartości miar przyjętych do oceny dokładności rozwiązania. Kolumna przedostatnia zawiera normę różnicy kolejnych przybliżeń, ostatnia zaś wartość względnego residuum wyznaczonego zgodnie z zależnością (7.13). Jak widać, w zależności od tego, jaka miara dokładności rozwiązania zostanie przyjęta, konieczne jest określenie odpowiedniej wartości oczekiwanej dokładności rozwiązania ε. Dla danego przybliżenia rozwiązania względna wartość residuum przyjmuje mniejsze wartości niż norma różnicy kolejnych przybliżeń. To powoduje, że można nieco wcześniej uznać, że wyznaczono rozwiązanie układu równań. Wynika to po części z tego, że kryterium (7.13) ocenia jakość bieżącego rozwiązania, zaś norma różnicy przybliżeń ocenia wielkość zmiany rozwiązania.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 7.2. Kolejne przybliżenia rozwiązania wraz z wartościami warunków skończenia obliczeń

i	$\mathbb{x}_i^T$				$\|\|\mathbb{x}_i - \mathbb{x}_{i-1}\|\|$	$\frac{\|\|\mathbb{A}\mathbb{x}_i - \mathbb{b}\|\|}{\|\|\mathbb{b}\|\|}$
	x_0	x_1	x_2	x_3		
0	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000
1	−0,5000	2,1250	0,7188	2,9844	3,7668	0,2581
2	−0,1055	1,8369	0,9705	2,9484	0,5507	0,0514
3	0,0205	2,0032	0,9918	2,9936	0,2145	0,0076
4	−0,0045	1,9964	0,9989	2,9990	0,0274	0,0015
5	$3{,}57 \cdot 10^{-4}$	1,9998	0,9998	2,9997	0,0060	0,000148
⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	$-3{,}72 \cdot 10^{-15}$	2,0000	1,0000	3,0000	$7{,}61 \cdot 10^{-14}$	$3{,}54 \cdot 10^{-15}$
20	$-6{,}11 \cdot 10^{-16}$	2,0000	1,0000	3,0000	$1{,}35 \cdot 10^{-14}$	$6{,}29 \cdot 10^{-16}$
21	$-1{,}11 \cdot 10^{-16}$	2,0000	1,0000	3,0000	$2{,}53 \cdot 10^{-15}$	$1{,}95 \cdot 10^{-16}$
22	0,0000	2,0000	1,0000	3,0000	$5{,}0877 \cdot 10^{-16}$	0,0000
23	0	2	1	3	0	0

A zatem przy normie różnicy kolejnych przybliżeń trzeba zrobić o jeden krok więcej, niż jest to konieczne, aby uzyskać informację o tym, że już osiągnięto wymaganą dokładność.

Jako kryterium zakończenia obliczeń przyjęto normę różnicy kolejnych przybliżeń rozwiązania. Po dwudziestu trzech iteracjach rozwiązanie osiąga dokładność maszynową. Widać zbieżność rozwiązania do wartości $\mathbb{x} = [0\,2\,1\,3]$. Jak łatwo sprawdzić, uzyskany wynik jest dokładnym rozwiązaniem układu równań (7.12).

Warunki zakończenia obliczeń są analogiczne do warunków dotyczących rozwiązywania równań algebraicznych w podrozdziale 3.9. Zazwyczaj jako warunek skończenia obliczeń wybiera się następującą zależność określającą błąd względny rozwiązania

$$\frac{||\mathbb{A}\mathbb{x}_i - \mathbb{b}||}{||\mathbb{b}||} \leqslant \varepsilon \tag{7.13}$$

gdzie ε jest wymaganym błędem względnym rozwiązania. Jak już wspomniano, z numerycznego punktu widzenia sposób wybrania normy nie jest aż tak istotny.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

7.5. Metody najszybszego spadku i gradientu sprzężonego

Rozwiązanie układu równań liniowych (7.14) można przeprowadzić, stosując pewną abstrakcyjną koncepcję. Zostanie ona w zarysie przedstawiona poniżej. Zatem, gdy dany jest układ równań zapisany macierzowo

$$\mathbb{A}\mathbb{x} = \mathbb{b} \tag{7.14}$$

gdzie $\mathbb{A}$ jest symetryczną dodatnio określoną macierzą o wymiarach $N \times N$, $\mathbb{x}$ ma wymiar $N \times 1$ i jest wektorem niewiadomych, zaś $\mathbb{b}$, o wymiarach $N \times 1$, jest wektorem prawej strony układu. W niektórych metodach numerycznych wymaganie dotyczące symetryczności macierzy często jest spełnione, gdy jest spełniona zasada wzajemności, która mówi o tym, że przy zamianie miejsc pobudzenia i pomiaru uzyska się identyczny wynik.

Po lewostronnym pomnożeniu układu równań przez $\mathbb{x}^T$ i wprowadzeniu czynnika skalującego można zdefiniować skalarną funkcję potencjału

$$\Pi = \frac{1}{2}\mathbb{x}^T\mathbb{A}\mathbb{x} - \mathbb{x}^T\mathbb{b} \tag{7.15}$$

Tak zdefiniowany potencjał Π przypisuje każdemu wektorowi $\mathbb{x}$ skalarną wartość, co pozwala na zdefiniowanie problemu rozwiązania układu równań (7.14) w postaci następującej pochodnej

$$\frac{\partial \Pi}{\partial \mathbb{x}} = 0 \Rightarrow \mathbb{A}\mathbb{x} - \mathbb{b} = 0 \tag{7.16}$$

Rodzina metod bazująca na minimalizacji tak określonego potencjału nazywana jest rodziną metod Kryłowa. W tych metodach rozwiązanie jest osiągane zazwyczaj przy znacznie mniejszej liczbie operacji niż N^3, jak to jest przy metodach bezpośrednich z rodziny metod Gaussa.

7.5.1. Metoda najszybszego spadku

Idea, na której opiera się metoda najszybszego spadku, może być ujęta następująco: jeżeli dla obecnej wartości $\mathbb{x}$ gradient potencjału się nie zeruje, to należy podążać przeciwnie do zwrotu gradientu, by zero osiągnąć. Ujemny gradient potencjału (7.16) jest równy residuum i określony jako

$$\mathbb{r}^i = -\frac{\partial \Pi}{\partial \mathbb{x}} = \mathbb{b} - \mathbb{A}\mathbb{x}^i$$

kolejne iteracje wykonywane są jako

$$\mathbb{x}^{i+1} = \mathbb{x}^i + \lambda^i \mathbb{r}^i$$

Należy zatem wybrać taką wartość λ^i, by znaleźć globalne minimum potencjału Π. Można napisać

$$\Pi = \frac{1}{2}\mathbb{x}^{T,i}\mathbb{A}\mathbb{x}^i + \lambda^i \mathbb{x}^{T,i}\mathbb{A}\mathbb{r}^i + \frac{1}{2}(\lambda^i)^2 \mathbb{r}^{T,i}\mathbb{A}\mathbb{r}^i - \mathbb{x}^{T,i}\mathbb{b} - \lambda^i \mathbb{r}^{T,i}\mathbb{b}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Pochodna potencjału Π jest równa

$$\frac{\partial \Pi}{\partial \lambda^i} = \mathbb{x}^{T,i}\mathbb{A}\mathbb{r}^i + \lambda^i \mathbb{r}^{T,i}\mathbb{A}\mathbb{r}^i - \mathbb{r}^{T,i}\mathbb{b} = 0$$

Jeżeli przyjmie się, że macierz $\mathbb{A}$ jest symetryczna (a jest to warunek stosowania metody), to można wyznaczyć wartość λ^i jako

$$\lambda^i = \frac{\mathbb{r}^{T,i}\left(\mathbb{b} - \mathbb{A}\mathbb{x}^i\right)}{\mathbb{r}^{T,i}\mathbb{A}\mathbb{r}^i} = \frac{\mathbb{r}^{T,i}\mathbb{r}^i}{\mathbb{r}^{T,i}\mathbb{A}\mathbb{r}^i}$$

Algorytm obliczeniowy 7.1 został przedstawiony poniżej. Sam algorytm nie jest skomplikowany. Szczególnie łatwo go zaimplementować w środowisku obliczeniowym wykonującym operacje na macierzach (np. MATLAB, GNU Octave).

Algorytm 7.1. Rozwiązywanie układu równań metodą najszybszego spadku

Require: $x_0 \neq 0, A, b, tol$
Ensure: A jest symetryczne
  repeat
    $r \leftarrow b - Ax_0$
    $\lambda \leftarrow r^T r/(r^T Ar)$
    $x_1 \leftarrow x_0 + \lambda r$
    $blad \leftarrow (x_1^T - x_0^T)A(x_1 - x_0)$
    $x_0 \leftarrow x_1$
  until $blad \geqslant tol$
  return x_0 rozwiązanie układu równań

Szybkość zbieżności algorytmu zależy od wskaźnika uwarunkowania macierzy współczynników $\mathbb{A}$ i oznaczając dokładne rozwiązanie jako $\mathbb{x}^*$, można napisać

$$||\mathbb{x}^* - \mathbb{x}^i|| = \left(1 - \frac{1}{C(\mathbb{A})}\right)^{\frac{i}{2}} ||\mathbb{x}^* - \mathbb{x}^1||$$

gdzie $C(\mathbb{A})$ jest ilorazem maksymalnej i minimalnej wartości własnej macierzy $\mathbb{A}$. Ta wersja algorytmu jest wolno zbieżna. Jej inny wariant – metoda gradientu sprzężonego osiąga wynik w najwyżej N iteracjach.

7.5.2. Metoda gradientu sprzężonego

Ponieważ idea metody bazuje na poszukiwaniu minimum w wybranych kierunkach, zaś kierunki poszukiwań są wyznaczane przez gradienty potencjału (7.15), metoda nosi nazwę gradientów sprzężonych (*Conjugate gradient*), w skrócie CG. Aby znaleźć przepis na iteracyjne rozwiązanie układu równań algebraicznych, podobnie jak poprzednio określa się residuum dla i-tej iteracji jako

$$\mathbb{r}^i = -\frac{\partial \Pi}{\partial \mathbb{x}} = \mathbb{b} - \mathbb{A}\mathbb{x}^i$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

zaś kolejne iteracje wykonywane są jako

$$\mathbb{x}^{i+1} = \mathbb{x}^i + \lambda^i \mathbb{z}^i$$

Poszukiwana jest taka wartość λ^i, aby potencjał Π osiągał minimum globalne. Pozostaje pytanie, jak wyznaczyć wektor $\mathbb{z}^i$. Wektor ten zależy od wielkości residuów w następujący sposób

$$\mathbb{z}^i = -\mathbb{r}^i + \beta^i \mathbb{z}^{i-1}$$

Współczynnik β określa, jaka część poprzedniego kierunku poszukiwania rozwiązania ma być użyta w bieżącej iteracji. Można go wyznaczyć z zależności

$$\beta^i = \frac{\mathbb{r}^{i,T}\mathbb{r}^i}{\mathbb{r}^{i-1,T}\mathbb{r}^{i-1}}$$

Po wyznaczeniu kierunku poszukiwań należy znaleźć minimum potencjału Π. Aby znaleźć wartość λ^i, można skorzystać z warunku zerowania się pochodnej potencjału Π, który na podstawie zależności (7.15) można zapisać jako

$$\begin{aligned}\Pi(\mathbb{x}^{i+1}) = \Pi(\mathbb{x}^i + \lambda^i \mathbb{x}^i) = \frac{1}{2}\left(\mathbb{x}^i + \lambda^i \mathbb{z}^i\right)^T \mathbb{A}\left(\mathbb{x}^i + \lambda^i \mathbb{z}^i\right) - \left(\mathbb{x}^i + \lambda^i \mathbb{z}^i\right)^T \mathbb{b} = \\ = \frac{1}{2}\mathbb{x}^{T,i}\mathbb{A}\mathbb{x}^i + \frac{1}{2}\lambda^i \mathbb{z}^{T,i}\mathbb{A}\mathbb{x}^i + \frac{1}{2}\lambda^i \mathbb{x}^{T,i}\mathbb{A}\mathbb{z}^i + +\frac{1}{2}(\lambda^i)^2 \mathbb{z}^{T,i}\mathbb{A}\mathbb{z}^i - \mathbb{x}^{T,i}\mathbb{b} - \lambda^i \mathbb{z}^{T,i}\mathbb{b}\end{aligned} \tag{7.17}$$

Pochodną potencjału można wyrazić jako

$$\frac{\partial \Pi}{\partial \lambda^i} = \frac{1}{2}\mathbb{z}^{T,i}\mathbb{A}\mathbb{x}^i + \frac{1}{2}\mathbb{x}^{T,i}\mathbb{A}\mathbb{z}^i + \lambda^i \mathbb{z}^{T,i}\mathbb{A}\mathbb{z}^i - \mathbb{z}^{T,i}\mathbb{b} = 0$$

Ponieważ macierz $\mathbb{A}$ jest symetryczna, pierwsze dwa składniki można połączyć i przenieść na prawą stronę równania, co da

$$\lambda^i \mathbb{z}^{T,i}\mathbb{A}\mathbb{z}^i = \mathbb{z}^{T,i}\mathbb{b} - \mathbb{z}^{T,i}\mathbb{A}\mathbb{x}^i = -\mathbb{z}^{T,i}\left(\mathbb{A}\mathbb{x}^i - \mathbb{b}\right) = -\mathbb{z}^{T,i}\mathbb{r}^i$$

Można zatem wyznaczyć wartość współczynnika λ^i jako

$$\lambda^i = -\frac{\mathbb{z}^{T,i}\mathbb{r}^i}{\mathbb{z}^{T,i}\mathbb{A}\mathbb{z}^i}$$

Współczynnik β zazwyczaj przyjmuje bardzo małe wartości, co powoduje, że $\mathbb{z}^{T,i} \approx -\mathbb{r}^{T,i}$, czyli dla zwiększenia efektywności obliczeń można przyjąć

$$\lambda^i = \frac{\mathbb{r}^{T,i}\mathbb{r}^i}{\mathbb{z}^{T,i}\mathbb{A}\mathbb{z}^i}$$

Jak widać, idea metody jest jasna i intuicyjna. Algorytm metody jest analogiczny do metody najszybszego spadku. W ramach ćwiczenia czytelnik może sam go opracować.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

7.5.3. Przyspieszanie obliczeń

Szybkość zbieżności metody gradientów sprzężonych silnie zależy od wskaźnika uwarunkowania macierzy $\mathbb{A}$. Zatem dla przyspieszenia zbieżności przekształca się wstępnie macierz $\mathbb{A}$ przez przekształcenie wstępne (*preconditioning*). Polega ono na tym, że zamiast rozwiązywać równanie $\mathbb{A}\mathbb{x} = \mathbb{b}$, rozwiązuje się równanie

$$\mathbb{P}^{-1}\mathbb{A}\mathbb{x} = \mathbb{P}^{-1}\mathbb{b}.$$

Gdy macierz $\mathbb{P}$ jest równa macierzy $\mathbb{A}$, to układ jest rozwiązany. Ważne jest, by iloczyn macierzy $\mathbb{P}^{-1}\mathbb{A}$ dawał macierz symetryczną. Ponieważ w ogólnym przypadku tak być nie musi, należy tak wybrać macierz $\mathbb{P}$, by można było ją rozłożyć na dwa czynniki $\mathbb{P} = \mathbb{E}\mathbb{E}^T$, co powoduje, że macierz $\mathbb{P}^{-1}$ wyraża się wzorem

$$\mathbb{P}^{-1} = \mathbb{E}^{-T}\mathbb{E}^{-1}$$

Niech wektor niewiadomych zostanie przekształcony do postaci

$$\mathbb{E}^T\mathbb{x} = \mathbb{y}, \quad \text{co oznacza, że} \quad \mathbb{x} = \mathbb{E}^{-T}\mathbb{y}$$

Po przyjęciu takich przekształceń można zapisać rozwiązywany układ równań jako

$$\mathbb{E}^{-T}\mathbb{E}^{-1}\mathbb{A}\mathbb{E}^{-T}\mathbb{y} = \mathbb{E}^{-T}\mathbb{E}^{-1}\mathbb{b}$$

Po koniecznych uproszczeniach można uzyskać następującą postać układu równań

$$\underbrace{\mathbb{E}^{-1}\mathbb{A}\mathbb{E}^{-T}}_{\widehat{\mathbb{A}}}\mathbb{y} = \underbrace{\mathbb{E}^{-1}\mathbb{b}}_{\widehat{\mathbb{b}}}$$

Czyli rozwiązywany będzie następujący układ równań

$$\widehat{\mathbb{A}}\mathbb{y} = \widehat{\mathbb{b}}$$

Nowa macierz $\widehat{\mathbb{A}}$ jest symetryczna i ma znacznie lepszy wskaźnik uwarunkowania, zaś cały układ równań można znacznie szybciej rozwiązać.

Jeszcze kilka słów na temat wyboru macierzy $\mathbb{P}^{-1}$. Macierz $\mathbb{P}$ powinna być możliwie bliska macierzy $\mathbb{A}$. Ale jednocześnie powinna być łatwa do odwracania. Jednym ze sposobów wyboru tej macierzy jest wybranie głównej przekątnej macierzy $\mathbb{A}$ jako macierzy $\mathbb{P}$. Jest to tak zwane wstępne przekształcenie Jacobiego (*Jacobi preconditioning*). Można też wybrać jedną, dwie lub trzy przekątne obok głównej. Dokonując rozkładu takiej macierzy, uzyskuje się metodę zwaną niekompletnym rozkładem Choleskiego (*incomplete Choleski decomposition*). W ten sposób uzyska się metodę określaną w skrócie ILU, odpowiednio 0, 1, 2 lub 3.

Zaletą tego sposobu poprawiania zbieżności metod Kryłowa jest to, że całe wstępne przekształcenie macierzy $\mathbb{A}$ można ukryć w algorytmie i nie zachodzi potrzeba jawnego wyznaczania rozkładu macierzy $\mathbb{P}$. Pomimo zwiększonego nakładu obliczeniowego związanego z wstępnym przekształcaniem, zysk na szybkości zbieżności metody jest tak duży, że warto stosować tego typu poprawianie uwarunkowania macierzy w układach równań liniowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

7.6. Podsumowanie

Metody iteracyjne są ekonomicznie uzasadnionym wyborem przy rozwiązywaniu układów o bardzo dużej liczbie niewiadomych. Nowoczesne metody rozwiązywania układów równań z symetrycznymi, dodatnio określonymi macierzami współczynników wymagają tylko N^2 operacji (np. metoda gradientu sprzężonego – CG). W porównaniu z eliminacją Gaussa, która wymaga około N^3, czy z metodą Cramera, która wymaga $2(N+1)!$ działań, oznacza znaczne przyspieszenie.

Tabela 7.3. Porównanie cech omawianych metod

Metoda	Znaczenie	Zalety	Wady	Zastosowanie
Eliminacja Gaussa	daje rozkład LU	prostota	duża złożoność obliczeniowa dla wielu podobnych układów	pojedyncze układy równań
Rozkład LU	duże	mała złożoność obliczeniowa	możliwa niestabilność dla źle uwarunkowanych układów równań	układy dobrze uwarunkowane
Rozkład QR	duże	stabilność numeryczna	większa złożoność obliczeniowa niż LU	układy źle uwarunkowane
CG	duże	nie wymaga znajomości parametrów układu	ograniczenie do szczególnej klasy równań	układy o macierzach symetrycznych, dodatnio określonych

Tabela 7.3 zawiera porównanie niektórych z omówionych do tej pory metod rozwiązywania układów równań [28]. W rozdziale przedstawiono główne grupy metod rozwiązywania układów liniowych równań algebraicznych. Ujęto jedynie niewielką część klasycznych metod stosowanych w istniejących systemach obliczeniowych. Czytelnik zainteresowany tematem może znaleźć obfitość wiedzy w literaturze podanej na końcu podręcznika.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 8

Aproksymacja

Można powiedzieć, że aproksymacja funkcji oznacza znajdowanie jej opisu matematycznego za pomocą funkcji aproksymującej. Funkcja aproksymowana może być wyrażona w postaci formuły lub podana jako tablica par argumentów i wartości.

W praktyce inżynierskiej często dochodzi do konieczności wykorzystania aproksymacji. Można chociażby wymienić: określanie współrzędnych środka otworu i jego promienia na podstawie współrzędnych kilku czy kilkunastu punktów na jego obwodzie, określenie wartości poziomu wielkości mierzonej w układzie mikroprocesorowym, co sprowadza się do wyznaczenia wartości funkcji $y = \log_{10} x$.

Jak wspomniano wcześniej, problemy aproksymacji spotykane w rzeczywistości można podzielić na dwie kategorie: problem opisu funkcji stablicowanej oraz problem polegający na zmianie opisu analitycznego funkcji danej w sposób jawny. Pierwszy z problemów może dotyczyć pomiarów, gdy tablica z wartościami funkcji powstaje w efekcie przeprowadzonych pomiarów. Jako przykład można podać określanie wartości oporu elektrycznego na podstawie pomiaru wartości spadku napięcia na oporniku przy zmianach przepływającego przez niego prądu. Zastosowanie aproksymacji do zagadnień dotyczących pomiarów parametrów obiektów rzeczywistych staje się bardzo istotne, gdy weźmie się pod uwagę fakt, iż nie ma możliwości wykonania pomiaru nieobarczonego błędem. Dlatego wyniki pomiarów dopasowywane są do przewidywań teoretycznych właśnie przez minimalizację błędu średniokwadratowego, co w istocie stanowi jedną z odmian aproksymacji.

Przy dokonywaniu aproksymacji funkcji $f(x)$ konieczne staje się rozwiązanie dwóch ważnych problemów. Pierwszym z nich jest dobór odpowiedniej funkcji aproksymującej $Q(x)$. Najczęściej jest to tzw. wielomian uogólniony stanowiący kombinację liniową funkcji bazowych $q_j(x)$

$$Q_m(x) = \sum_{j=0}^{m} a_j q_j(x)$$

Jako funkcje bazowe stosowane bywają między innymi: jednomiany: 1, x, x^2, ...,x^m, funkcje trygonometryczne: 1, $\sin x$, $\cos x$, $\sin 2x$, $\cos 2x$, ..., $\sin mx$, $\cos mx$, oraz wielomiany ortogonalne. Przyjęcie odpowiednich funkcji bazowych powoduje, że problem aproksymacji sprowadza się do wyznaczenia wartości współczynników a_0, a_1, ..., a_m, co stanowi drugi problem, który przy aproksymacji należy rozwiązać.

Oczywiste jest, że aproksymacja funkcji innymi funkcjami powoduje powstanie błędów. Aproksymacja jest procesem optymalizacyjnym prowadzącym do minimalizacji tych błędów, zaś sposób ich określania decyduje o wyborze metody aproksymacji. Jeśli błąd będzie określany dla dyskretnego zbioru węzłów x_0, x_1,..., x_n, to jest to

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

aproksymacja punktowa. Z kolei gdy miarą błędu będzie różnica funkcji w pewnym przedziale $\langle a, b\rangle$, to będzie to aproksymacja integralna lub przedziałowa. Zazwyczaj ten drugi przypadek zachodzi, gdy znana jest analityczna postać funkcji aproksymowanej.

Najczęściej stosowane sposoby określania błędów podczas aproksymacji zostały zestawione poniżej:

— dla aproksymacji średniokwadratowej punktowej

$$S = \sum_{i=0}^{n} \left[f(x_i) - Q(x_i)\right]^2,$$

— dla aproksymacji średniokwadratowej integralnej lub przedziałowej

$$S = \int_a^b \left[f(x) - Q(x)\right]^2 \mathrm{d}x,$$

— dla aproksymacji jednostajnej

$$S = \sup_{x\in[a,b]} |f(x) - Q(x)|.$$

We wszystkich tych przypadkach zadanie aproksymacji sprowadza się do takiego doboru funkcji aproksymującej, aby uzyskać minimalizację tak określonej funkcji błędów. Zatem problem aproksymacji można rozumieć jako problem optymalizacyjny. Poszukuje się takiego zestawu współczynników, by funkcja celu osiągała wartość minimalną.

8.1. Aproksymacja punktowa wielomianowa

Zadanie aproksymacji punktowej wielomianowej można postawić następująco: znaleźć wielomian $p(x)$ wybranego stopnia taki, że dla n punktów (x_i, y_i) zachodzi

$$\min ||p(x_i) - y_i||, \quad \text{dla} \quad i = 0, 1, 2, \ldots$$

W zależności od wyboru normy $||\cdot||$, którą można rozumieć jako miarę odległości między zadanymi punktami i wartościami funkcji aproksymującej, można mówić o różnych rodzajach aproksymacji, np. średniokwadratowej czy optymalnej.

8.1.1. Aproksymacja średniokwadratowa

Jeżeli za miarę jakości dopasowania funkcji, którą w omawianym przypadku stanowi wielomian stopnia k, do zbioru punktów przyjmie się sumę kwadratów odchyłek wartości funkcji aproksymującej od zadanych wartości, to aproksymacja nazywa się średniokwadratową. Oczywiście można dokonywać aproksymacji średniokwadratowej

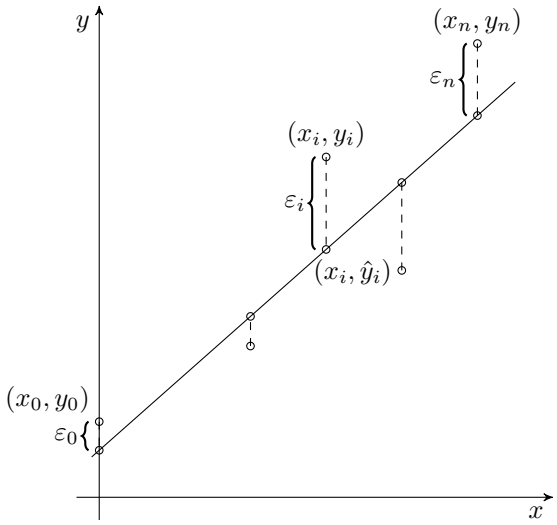
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

za pomocą innych funkcji, ale nie zawsze da się wyprowadzić tak eleganckie zależności, jak dla aproksymacji wielomianowej. Jednak okazuje się, że wiele zależności w technice można dostatecznie dokładnie opisać wielomianem stosunkowo niskiego stopnia, gdy się je przedstawi w skali logarytmicznej. Stąd też aproksymacja wielomianowa stanowi tak przydatne narzędzie dla inżyniera i jest tak rozpowszechniona.

Zadanie aproksymacji formułuje się następująco:
Wyznaczyć wartości współczynników wielomianu $W(x)$ stopnia k, który dla danego zbioru węzłów aproksymacji x_i z wartościami y_i, gdzie $i = 0, 1, .., n$, przy czym $k < n$, przyjmie takie wartości $\hat{y}_i$, że spełniony zostanie warunek

$$\min\left(\sum_i (\hat{y}_i - y_i)^2\right) \tag{8.1}$$

Rysunek 8.1 poglądowo wyjaśnia ideę aproksymacji punktowej wielomianowej. Przedstawiono na nim punkty stanowiące aproksymowaną funkcję, wynikowy wielomian aproksymujący oraz zaznaczono błędy ε_i, których suma kwadratów ma być minimalizowana w trakcie aproksymacji.



Rysunek 8.1. Idea zagadnienia aproksymacji

Przyjmuje się, że wielomian $W(x)$ jest stopnia k, i ma postać (8.2). Przyjęta konwencja zapisu wielomianu ma znaczenie. Jeżeli zapis zacznie się od współczynnika przy najwyższej potędze, dalsze rozważania przyjmą nieco inną postać.

$$W(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k \tag{8.2}$$

Podobnie jak w interpolacji, aby rozwiązać tak postawione zadanie, wymaga się, by wartości wielomianu aproksymującego $W(x)$ w kolejnych węzłach x_i wynosiły y_i. Takie rozumowanie prowadzi do układu równań (8.3), który jest nadokreślony, czyli

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

zawiera więcej równań niż niewiadomych współczynników oraz w ogólnym przypadku jest sprzeczny.

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_k x_0^k &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_k x_1^k &= y_1 \\ a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_k x_2^k &= y_2 \\ \qquad\qquad \vdots & \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_k x_n^k &= y_n \end{cases} \tag{8.3}$$

Ponieważ rząd macierzy współczynników wynosi najwyżej $k+1$, powyższy układ równań nie może być w tej postaci rozwiązany metodami służącymi do rozwiązywania układów algebraicznych równań liniowych.

Układ równań (8.3) można zwięźle zapisać w postaci macierzowej

$$\mathbb{X}\mathbb{A} = \mathbb{Y} \tag{8.4}$$

gdzie macierz

$$\mathbb{X} = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^k \\ 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{bmatrix} \tag{8.5}$$

ma znaną z interpolacji postać macierzy Vandermonde'a. Oczywiste jest, że rozmiary macierzy $\mathbb{X}$ zależą od liczby węzłów aproksymacji oraz od stopnia wielomianu aproksymującego. Macierz $\mathbb{A}$ jest macierzą kolumnową nieznanych współczynników wielomianu aproksymacyjnego i ma postać zależną od sposobu zapisu tego wielomianu we wzorze (8.2).

$$\mathbb{A} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Gdy wielomian aproksymacyjny, opisany zależnością (8.2), zostanie zapisany w odwrotnej kolejności, macierz $\mathbb{X}$ przyjmie postać będącą lustrzanym odbiciem macierzy z równania (8.5). Odwróceniu ulegnie też kolejność współczynników wielomianu stanowiących elementy macierzy $\mathbb{A}$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wartości wielomianu aproksymującego $\hat{\mathbb{Y}}$ można wyznaczyć jako $\mathbb{X}\mathbb{A} = \hat{\mathbb{Y}}$, zaś różnica $\hat{y}_i - y_i$, występująca pod znakiem sumy w zależności (8.1) zostanie oznaczona przez ε_i. Wykorzystując wprowadzone oznaczenia, sumę kwadratów błędów występującą w zależności (8.1) można zapisać jako

$$S = \sum_{i=0}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=0}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = \left(\hat{\mathbb{Y}} - \mathbb{Y}\right)^T \left(\hat{\mathbb{Y}} - \mathbb{Y}\right) \tag{8.6}$$

Ponieważ $\mathbb{X}\mathbb{A} = \hat{\mathbb{Y}}$, można napisać

$$S = \sum_{i=0}^{n} \varepsilon_i^2 = (\mathbb{X}\mathbb{A} - \mathbb{Y})^T (\mathbb{X}\mathbb{A} - \mathbb{Y}) = \mathbb{A}^T\mathbb{X}^T\mathbb{X}\mathbb{A} - 2\mathbb{A}^T\mathbb{X}^T\mathbb{Y} + \mathbb{Y}^T\mathbb{Y} \tag{8.7}$$

Minimum wyznaczonej sumy kwadratów błędów można określić z warunku zerowania się pochodnych

$$\frac{\partial S}{\partial \mathbb{A}} = 0, \quad \text{co można zapisać jako} \quad \begin{bmatrix} \frac{\partial S}{\partial a_0} \\ \frac{\partial S}{\partial a_1} \\ \frac{\partial S}{\partial a_2} \\ \vdots \\ \frac{\partial S}{\partial a_k} \end{bmatrix} = 0 \tag{8.8}$$

Pochodna wyrażenia (8.7) względem współczynników wielomianu wynosi

$$\frac{\partial S}{\partial \mathbb{A}} = \frac{\partial}{\partial \mathbb{A}} \left(\mathbb{A}^T\mathbb{X}^T\mathbb{X}\mathbb{A} - 2\mathbb{A}^T\mathbb{X}^T\mathbb{Y} + \mathbb{Y}^T\mathbb{Y}\right) = 2\mathbb{X}^T\mathbb{X}\mathbb{A} - 2\mathbb{X}^T\mathbb{Y}$$

Z powyższej zależności, po przyrównaniu do zera, otrzymuje się warunek (8.8), który można zapisać w postaci układu równań normalnych Gaussa (8.9).

$$\mathbb{X}^T\mathbb{X}\mathbb{A} = \mathbb{X}^T\mathbb{Y} \tag{8.9}$$

Układ równań normalnych ma jednoznaczne rozwiązanie, jeżeli liczba węzłów $n+1$ jest większa od liczby współczynników $k+1$ lub jej równa. Niestety, dla $k \geqslant 5$ aproksymacja wielomianowa staje się źle uwarunkowana i zalecane jest wykorzystywanie wielomianów ortogonalnych.

Proste przekształcenie powyższej zależności pozwala na określenie optymalnego, w sensie średniokwadratowym, wektora współczynników wielomianu $\mathbb{A}$ jako

$$\mathbb{A} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1} \mathbb{X}^T\mathbb{Y} \tag{8.10}$$

Nie można dać się zwieść elegancji powyższej zależności. Wyniki uzyskiwane po jej zastosowaniu mogą być obarczone dużymi błędami, zwłaszcza jeśli iloczyn $\mathbb{X}^T\mathbb{X}$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

jest słabo uwarunkowany. Problemy pojawiają się głównie przy korzystaniu z systemów o niezbyt długim słowie maszynowym, co skutkuje stosunkowo niewielką liczbą cyfr znaczących. Dla poprawienia dokładności estymacji współczynników wielomianu aproksymującego czasami zaleca się stosowanie przekształcenia Householdera. Taki zabieg powoduje uzyskanie dwa razy większej liczby cyfr dokładnych, niemniej jest około dwa razy bardziej kosztowny obliczeniowo [37].

Inną metodą rozwiązania problemu średniokwadratowej aproksymacji wielomianowej jest skorzystanie z rozkładu względem wartości szczególnych (SVD) [23] i pseudoodwrotności Moore'a–Penrose'a. Wykorzystując liniowy układ równań algebraicznych (8.4) zapisany dla $n+1$ węzłów i wielomianu stopnia k

$$\mathbb{X}\mathbb{A} = \mathbb{Y}$$

gdzie macierz $\mathbb{X}$ ma $n+1$ wierszy oraz $k+1$ kolumn, można znaleźć rozwiązanie minimalne w postaci

$$\mathbb{A} = \mathbb{X}^{\dagger}\mathbb{Y}$$

Aby wyznaczyć pseudoodwrotną macierz $\mathbb{X}^{\dagger}$, można wykorzystać rozkład względem wartości osobliwych (SVD), dzięki któremu rozkłada się macierz $\mathbb{X}$ na trzy macierze

$$\mathbb{X} = \mathbb{P}\mathbb{D}\mathbb{Q}$$

gdzie $\mathbb{P}$ jest macierzą unitarną stopnia $n+1$, $\mathbb{Q}$ jest macierzą unitarną stopnia $k+1$, zaś $\mathbb{D}$ jest macierzą diagonalną o $n+1$ wierszach i $k+1$ kolumnach, a zawiera na przekątnej wartości osobliwe macierzy $\mathbb{X}$ oznaczane przez σ_i. Pseudoodwrotność macierzy $\mathbb{X}$ wyznacza się jako

$$\mathbb{X}^{\dagger} = \mathbb{Q}^{*}\mathbb{D}^{\dagger}\mathbb{P}^{*}$$

gdzie * oznacza sprzężenie macierzy. Macierz $\mathbb{D}^{\dagger}$ składającą się z $k+1$ wierszy i $n+1$ kolumn, oblicza się jako

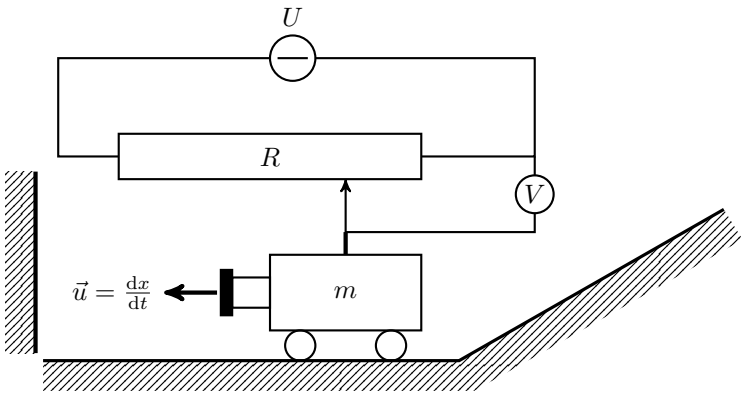
$$\mathbb{D}^{\dagger}_{ij} = \begin{cases} \sigma_i^{-1} & \text{dla } i = j < \operatorname{rank}(\mathbb{X}) \\ 0 & \text{w przeciwnym razie} \end{cases}$$

Wyniki uzyskiwane za pomocą pseudoodwrtotności są zgodne z wynikami uzyskiwanymi za pomocą zależności (8.10), ponieważ pseudoodwrotność Moore'a–Penrose'a stanowi efektywną numerycznie realizację metody najmniejszych kwadratów.

8.1.2. Przykład praktyczny

W układzie pomiarowym służącym do badania układów pochłaniających energię przy zderzeniu zastosowano potencjometr suwakowy R do określania położenia wózka w końcowym etapie ruchu. Na zaciskach wyjściowych dzielnika napięcia pojawia się napięcie elektryczne proporcjonalne do wielkości przemieszczenia wózka od początku potencjometru. Pomiar tego napięcia pozwala na określenie położenie wózka w każdej chwili czasowej. Za pomocą karty pomiarowej dokonywany jest pomiar napięcia z okresem T.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 8.2. Schemat stanowiska do badania układu pochłaniającego energię zderzeń

Układ pomiarowy z symbolicznie zaznaczonym połączeniem wózka z potencjometrem przedstawiono na rysunku 8.2. Zadanie polega na wyznaczeniu prędkości wózka przed zderzeniem, w celu określenia jego energii kinetycznej i skuteczności układu pochłaniania energii. Prędkość średnią należy wyznaczyć metodą aproksymacji. Przebieg aproksymować funkcją liniową. Współczynnik kierunkowy prostej regresji będzie poszukiwaną średnią wartością prędkości wózka. Należy zwrócić uwagę na jednostki, w których wynik będzie wyrażony. Wózek wraz z czujnikami i układem pochłaniającym energię zderzenia ma masę 2 kg. Okres próbkowania wynosi $T = 10^{-4}$ s. Sam moment zderzenia następuje w 243. próbce.

8.1.3. Rozwiązanie

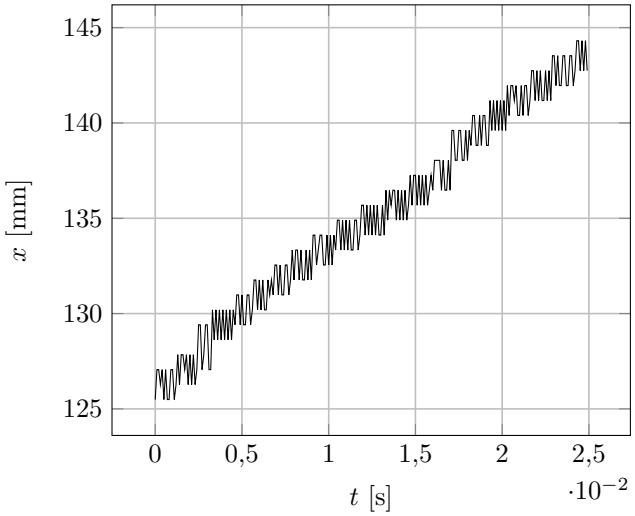
Zarejestrowany przebieg przemieszczeń w czasie przedstawia rysunek 8.3. Widoczny szum wynika z rozdzielczości układu pomiarowo-rejestrującego. Przemieszczenia wyznaczono, mnożąc zarejestrowane napięcia przez współczynnik uzyskany na etapie kalibracji stanowiska.

Opisywany problem aproksymacji można rozwiązać na dwa sposoby. Pierwszy z nich będzie polegał na wykorzystaniu zależności (8.10). Wiąże się to ze zbudowaniem macierzy $\mathbb{X}$ oraz $\mathbb{Y}$, a następnie wyznaczeniem rozwiązania w postaci wektora współczynników wielomianu pierwszego stopnia najlepiej, w sensie średniokwadratowym, opisującego zbiór aproksymowanych punktów.

Jako pierwszy krok rozwiązania problemu należy wybrać wielomian aproksymacyjny. Ponieważ poszukiwana jest średnia prędkość, więc w analizowanych przypadku będzie to wielomian pierwszego stopnia, czyli wielomian postaci

$$\hat{y} = a_0 + a_1 t$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 8.3. Przebieg zarejestrowanego przemieszczenia wózka w czasie

Należy ułożyć układ równań (8.3). Macierz współczynników wielomianu $\mathbb{A}$, macierz $\mathbb{X}$ oraz macierz wartości pomiarowych $\mathbb{Y}$ będą miały postać następującą

$$\mathbb{X} = \begin{bmatrix} 1 & 0 \\ 1 & 10^{-4} \\ 1 & 2 \cdot 10^{-4} \\ \vdots & \vdots \\ 1 & 2{,}42 \cdot 10^{-2} \end{bmatrix}, \qquad \mathbb{A} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}, \qquad \mathbb{Y} = \begin{bmatrix} 125{,}49 \\ 127{,}06 \\ 127{,}06 \\ 126{,}27 \\ \vdots \\ 141{,}96 \\ 142{,}75 \end{bmatrix}$$

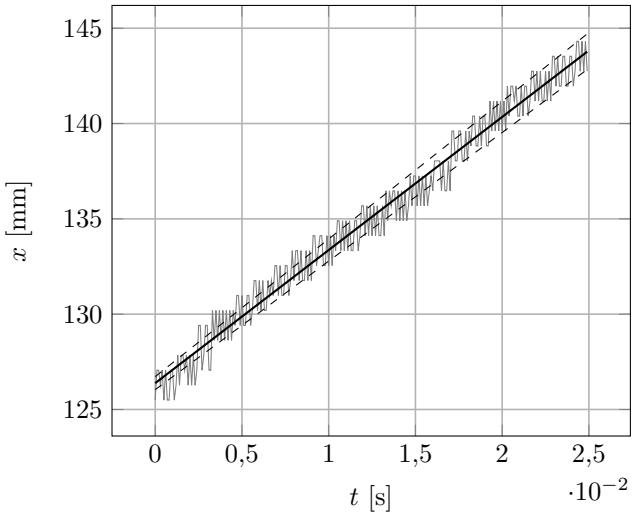
Ze względu na oszczędność miejsca zostały przedstawione jedynie fragmenty macierzy $\mathbb{X}$ oraz $\mathbb{Y}$. W rzeczywistości macierze te mają po 243 wiersze. Wyznaczenie współczynników zgodnie z wzorem (8.10) przedstawiono w równaniu (8.11), przy czym, podobnie jak poprzednio, macierze, które są zbyt duże, by je przedstawić na stronie, zostały skrócone.

$$\mathbb{A} = \left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \mathbb{Y} = \begin{bmatrix} 243{,}0 & 2{,}9403 \\ 2{,}9403 & 0{,}0475349 \end{bmatrix}^{-1}.$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\cdot \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 10^{-4} & 2 \cdot 10^{-4} & \dots & 2{,}42 \cdot 10^{-2} \end{bmatrix} \begin{bmatrix} 125{,}49 \\ 127{,}06 \\ 127{,}06 \\ \vdots \\ 141{,}96 \\ 142{,}75 \end{bmatrix} = \begin{bmatrix} 126{,}39 \\ 697{,}74 \end{bmatrix} \quad (8.11)$$

Drugi sposób polega na wykorzystaniu gotowych funkcji służących do aproksymacji, na przykład w MATLAB-ie oraz GNU Octave jest to `polyfit`.



Rysunek 8.4. Wyznaczona prosta regresji oraz przedziały ufności

Jako argumenty podaje się wartości zmiennych niezależnych i zależnych w punktach oraz stopień wielomianu. Zwracane są wartości współczynników wielomianu, który ma najmniejszą sumę kwadratów odchyłek od podanych punktów. Oprócz współczynników wielomianu minimalizującego błąd średniokwadratowy można uzyskać też informacje o jakości dopasowania i przedziałach ufności.

```
  [A,S]=polyfit(t,y,1);
2 odchA=sqrt(diag(S.C)/S.df)*S.normr
  yapr=polyval(A,t);
4 figure(1)
  plot(t,y,t,yapr,t,polyval(A+3*odchA',t),t,polyval(A-3*odchA',t))
6 title('Aproksymacja')
```

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Powyższy listing przedstawia rozwiązanie omawianego problemu. Zmienne `t` oraz `y` zawierają odpowiednio chwile czasowe oraz zmierzone wartości przemieszczenia. W wyniku uzyskuje się wykres taki jak na rysunku 8.4.

Wartość średniej prędkości wózka przed zderzeniem wynosi 697,74 mm/s. Problem można też rozwiązać za pomocą pseudoodwrotności Moore'a–Penrose'a, otrzymując analogiczne wyniki.

8.2. Podsumowanie

W niniejszym rozdziale przedstawiono zarys problematyki związanej z aproksymacją, a i to jedynie związanej z aproksymacją punktową wielomianową. Nieco więcej informacji na ten temat zostanie przedstawionych w rozdziale 13. Zostanie omówiona, między innymi, aproksymacja nieliniowa i jakość aproksymacji.

Celem niniejszego rozdziału jest podanie podstawowych informacji przydatnych dla inżyniera do obróbki danych pomiarowych. Wygładzanie czy odszumianie może być zrealizowane za pomocą podanych tu zależności. Co więcej te informacje pozwalają na samodzielną implementację filtrów Savitzkiego–Golaya. Autorzy rozmyślnie nie podali informacji dotyczących statystycznej oceny dopasowania wyników, wychodząc z założenia, że czytelnik najczęściej taką ocenę przeprowadza wizualnie.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Część II

Zaawansowane metody numeryczne

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 9

Metoda różnic skończonych

9.1. Wstęp

Historycznie rzecz biorąc, można powiedzieć, że metoda różnic skończonych jest najwcześniejszą metodą numerycznego rozwiązywania równań różniczkowych, szczególnie problemów brzegowych i początkowo-brzegowych. Istota metody różnic skończonych sprowadza się do dyskretyzacji analizowanego obszaru, a następnie zastąpieniu operatorów różniczkowania w równaniu odpowiednimi operatorami różnicowymi, określonymi na na dyskretnym zbiorze punktów. Ten dyskretny zbiór punktów nazywany jest siatką, a jego elementy nazywane są węzłami.

Chociaż sama metoda różnic skończonych nie wymaga regularnej siatki, to dla uproszczenia rozważań przyjęto siatkę regularną ze stałym krokiem.

Operatory różnicowe w przypadku problemu jednowymiarowego zostały omówione w rozdziale 5, dotyczącym różniczkowania numerycznego. Jak wspomniano, najprostszym sposobem przybliżenia wartości pochodnej w punkcie jest iloraz różnicowy. Ilorazy różnicowe buduje się następująco:

— iloraz różnicowy pierwszego rzędu w przód

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_i} = \frac{y_{i+1} - y_i}{h},$$

— iloraz różnicowy pierwszego rzędu w tył

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_i} = \frac{y_i - y_{i-1}}{h},$$

— iloraz różnicowy pierwszego rzędu centralny

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=x_i} = \frac{y_{i+1} - y_{i-1}}{h},$$

— iloraz różnicowy drugiego rzędu

$$\left.\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right|_{x=x_i} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}.$$

Oprócz ilorazów różnicowych pochodne można aproksymować bardziej złożonymi wyrażeniami.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

9.2. Pochodna pierwszego rzędu

Często używa się kilku punktów do przybliżania wartości pochodnej w punkcie, uzyskując w ten sposób lepszą dokładność. Na przykład pierwsze pochodne na siatce ze stałym krokiem można wyznaczać na podstawie zależności zawartych w tabeli 9.1. Jeden ze sposobów uzyskiwania tych zależności przedstawiono wzorami (5.19) do (5.22). Po przeprowadzeniu interpolacji wielomianowej różniczkuje się otrzymany wielomian, a następnie wyznacza się wartości pochodnych w poszczególnych punktach. Aby uzyskać informację o wielkości błędu, wykorzystuje się rozwinięcie funkcji w szereg Taylora i określa się wielkość błędu obcięcia tego szeregu dla danej interpolacji.

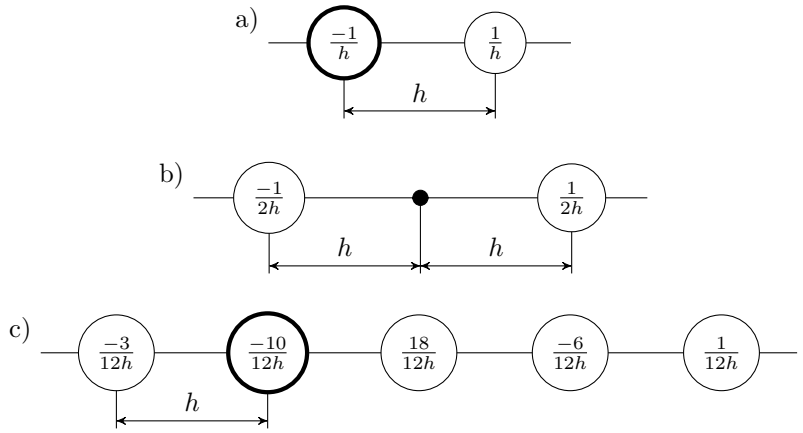
Tabela 9.1. Wzory różnicowe dla pierwszej pochodnej na siatce z węzłami równo odległymi

Lp.	Zależność	Uwagi
1	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{f_{i+1}-f_i}{h} + \mathcal{O}(h)$	dwu-punktowe
2	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{f_{i+1}-f_{i-1}}{2h} + \mathcal{O}(h^2)$	
3	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{f_i-f_{i-1}}{h} + \mathcal{O}(h)$	
4	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{2h}\left[-3f_i + 4f_{i+1} - f_{i+2}\right] + \mathcal{O}(h^2)$	trój-punktowe
5	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{2h}\left[f_{i-2} - 4f_{i-1} + 3f_i\right] + \mathcal{O}(h^2)$	
6	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{6h}\left[-11f_i + 18f_{i+1} - 9f_{i+2} + 2f_{i+3}\right] + \mathcal{O}(h^3)$	cztero-punktowe
7	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{6h}\left[-2f_{i-1} - 3f_i + 6f_{i+1} - f_{i+2}\right] + \mathcal{O}(h^3)$	
8	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{6h}\left[f_{i-2} - 6f_{i-1} + 3f_i + 2f_{i+1}\right] + \mathcal{O}(h^3)$	
9	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{6h}\left[-2f_{i-3} + 9f_{i-2} - 18f_{i-1} + 11f_i\right] + \mathcal{O}(h^3)$	
10	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{12h}[-25f_i + 48f_{i+1} - 36f_{i+2} + 16f_{i+3} - 3f_{i+4}] + \mathcal{O}(h^4)$	pięcio-punktowe
11	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{12h}[-3f_{i-1} - 10f_i + 18f_{i+1} - 6f_{i+2} + f_{i+3}] + \mathcal{O}(h^4)$	
12	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{12h}[f_{i-2} - 8f_{i-1} + 8f_{i+1} - f_{i+2}] + \mathcal{O}(h^4)$	
13	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{12h}[-f_{i-3} + 6f_{i-2} - 18f_{i-1} + 10f_i + 3f_{i+1}] + \mathcal{O}(h^4)$	
14	$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right\|_i = \frac{1}{12h}[3f_{i-4} - 16f_{i-3} + 36f_{i-2} - 48f_{i-1} + 25f_i] + \mathcal{O}(h^4)$	

Operatory różnicowe służące do przybliżonego wyznaczania pochodnej można przedstawiać graficznie. Taka graficzna reprezentacja operatora różnicowego nazywa się

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

schematem różnicowym lub gwiazdą. Przykładowe jednowymiarowe schematy różnicowe przedstawiono na rysunku 9.1. Węzły są zaznaczane okręgami; węzeł, w którym jest wyznaczana wartość pochodnej jest narysowany grubą linią. Pierwszy schemat różnicowy (rys. 9.1a) odwzorowuje pierwszy wiersz tabeli, zawierający iloraz różnicowy „w przód". Wartości wewnątrz okręgów oznaczają wagi, przez które są mnożone wartości w odpowiednich węzłach. Drugi schemat (rys. 9.1b) odpowiada ilorazowi różnicowemu centralnemu, zapisanemu w drugim wierszu tabeli 9.1.



Rysunek 9.1. Przykładowe schematy różnicowe dla wierszy: a) – 1, b) – 2 i c) – 11 tabeli 9.1

W przypadku ilorazu różnicowego centralnego węzeł, w którym jest wyznaczana wartość pochodnej, został zaznaczony czarną kropką. Ostatni schemat (rys. 9.1c) przedstawia równanie z 11 wiersza wspomnianej tabeli. Odległości między węzłami są stałe i wynoszą h. Widać, że schematy różnicowe jednoznacznie przedstawiają równania różnicowe pochodnych w poszczególnych węzłach.

9.3. Pochodne wyższych rzędów

Przybliżanie pochodnych wyższych rzędów odbywa się w sposób analogiczny do przedstawionego powyżej. Na przykład, wykorzystując rozwinięcie w szereg Taylora w przód i w tył, można wyprowadzić zależność różnicową na drugą pochodną.

$$f(x_i+h) = f(x_i) + hf'(x_i) + \frac{h^2}{2!}f''(x_i) + \frac{h^3}{3!}f'''(x_i) + \cdots \quad (9.1)$$

$$+$$

$$f(x_i-h) = f(x_i) - hf'(x_i) + \frac{h^2}{2!}f''(x_i) - \frac{h^3}{3!}f'''(x_i) + \cdots \quad (9.2)$$

$$f(x_i+h) + f(x_i-h) = 2f(x_i) + 2\frac{h^2}{2!}f''(x_i) + 2\frac{h^4}{4!}f^{(4)}(x_i) + \cdots \quad (9.3)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Dodanie stronami równań (9.1) oraz (9.2) daje równanie (9.3), które po obcięciu do drugiej pochodnej można zapisać w postaci

$$f(x_i + h) + f(x_i - h) = 2f(x_i) + h^2 f''(x_i) + \mathcal{O}(h^4)$$

Zatem druga pochodna może być aproksymowana następującym równaniem różnicowym

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{f(x_i + h) - 2f(x_i) + f(x_i - h)}{h^2} + \mathcal{O}(h^2)$$

które można zapisać jako

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} + \mathcal{O}(h^2) \tag{9.4}$$

Innym sposobem wyznaczenia przybliżeń drugiej pochodnej jest wykorzystanie definicji drugiej pochodnej, która mówi, że jest to pochodna z pierwszej pochodnej. Wykorzystując odpowiednie wyrażenia różnicowe, można skonstruować wyrażenia na drugą pochodną.

Jeszcze inny sposób polega na znalezieniu wielomianu interpolacyjnego dla zbioru punktów, wyznaczenie drugiej pochodnej i określania jej wartości w poszczególnych węzłach. Na przykład, dla gwiazdy budowanej z pięciu punktów, drugie pochodne w poszczególnych punktach będą się wyrażać poniższymi zależnościami.

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{1}{12h^2}\left[35\, f_i - 104\, f_{i+1} + 114\, f_{i+2} - 56\, f_{i+3} + 11\, f_{i+4}\right]$$

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{1}{12h^2}\left[11\, f_{i-1} - 20\, f_i + 6\, f_{i+1} + 4\, f_{i+2} - f_{i+3}\right]$$

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{1}{12h^2}\left[-f_{i-2} + 16\, f_{i-1} - 30\, f_i + 16\, f_{i+1} - f_{i+2}\right]$$

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{1}{12h^2}\left[-f_{i-3} + 4\, f_{i-2} + 6\, f_{i-1} - 20\, f_i + 11\, f_{i+1}\right]$$

$$\left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_i = \frac{1}{12h^2}\left[+11\, f_{i-4} - 56\, f_{i-3} + 114\, f_{i-2} - 104\, f_{i-1} + 35\, f_i\right]$$

W podobny sposób można wyznaczyć wyrażenia różnicowe dla pochodnych wyższych rzędów, czy dla innej liczby punktów, które są wykorzystywane w aproksymacji równania różniczkowego.

9.4. Pochodne mieszane

Oprócz problemów jednowymiarowych można rozwiązywać zagadnienia z większą liczbą zmiennych niezależnych. Wymaga to przybliżonego określania pochodnych mie-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

szanych. Postępowanie przy wyrażaniu pochodnych mieszanych za pomocą ilorazów różnicowych wynika wprost z definicji pochodnej mieszanej.

$$\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x_i,y_j} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)_{x_i,y_j} \tag{9.5}$$

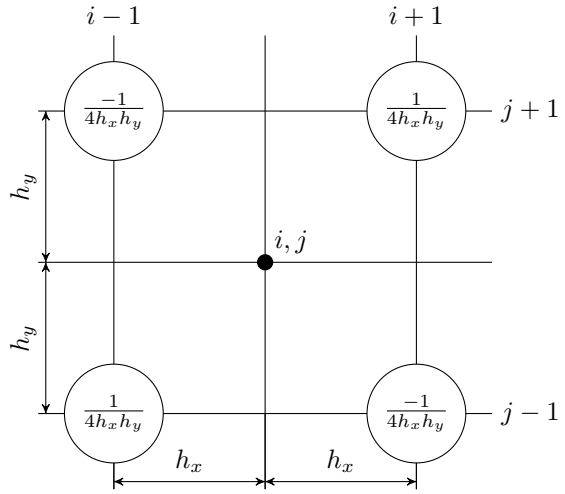
Zatem podobnie jak pochodne pierwszego rzędu można pochodne mieszane przybliżyć wyrażeniami różnicowymi. Co więcej, pochodne względem poszczególnych zmiennych można wyrażać przybliżeniami różnych rzędów, uzyskując wyrażenia dowolnej dokładności. Na przykład przybliżenie mieszanej pochodnej za pomocą centralnych ilorazów różnicowych będzie wymagać aproksymowania pochodnej względem zmiennej x

$$\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x_i,y_j} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right)_{x_i,y_j} = \frac{1}{2h_x}\left(\left.\frac{\partial f}{\partial y}\right|_{i+1,j} - \left.\frac{\partial f}{\partial y}\right|_{i-1,j}\right) + \mathcal{O}(h_x^2)$$

a następnie aproksymowania pochodnej względem zmiennej y. Co można zapisać jako

$$\begin{aligned}\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{x_i,y_j} &= \frac{1}{2h_x}\left(\frac{f_{i+1,j+1} - f_{i+1,j-1}}{2h_y} - \frac{f_{i-1,j+1} - f_{i-1,j-1}}{2h_y}\right) + \mathcal{O}(h_x^2, h_y^2) = \\ &= \frac{1}{4h_x h_y}\left(f_{i+1,j+1} - f_{i+1,j-1} - f_{i-1,j+1} + f_{i-1,j-1}\right) + \mathcal{O}(h_x^2, h_y^2)\end{aligned} \tag{9.6}$$

Rysunek 9.2 przedstawia schemat różnicowy ilustrujący zależność (9.6). Zaczerniona kropka oznacza węzeł, w którym wyznacza się wartość pochodnej, wartości wewnątrz okręgów oznaczających otaczające węzły są współczynnikami przy wartościach poszukiwanej funkcji w poszczególnych węzłach.



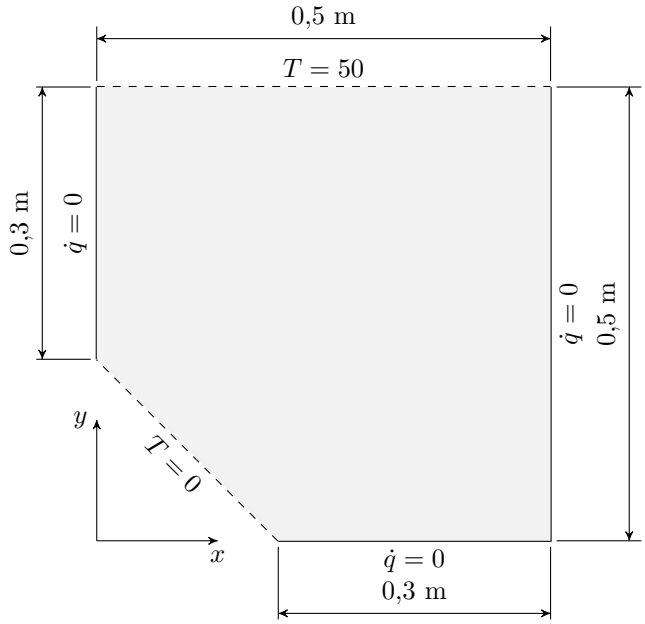
Rysunek 9.2. Schemat różnicowy dla zależności (9.6)

Jak widać, schematy różnicowe są bardzo intuicyjnymi sposobami ilustrowania równań różnicowych. Najlepiej jest to widoczne przy regularnych siatkach węzłów.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

9.5. Przykład dwuwymiarowej analizy

Sformułowanie problemu. Wyznaczyć rozkład temperatury w stanie ustalonym dla obszaru przedstawionego na rysunku 9.3. Wykorzystać schematy zapewniające dokładność rzędu $\mathcal{O}(h^2)$.



Rysunek 9.3. Domena obliczeniowa z warunkami brzegowymi

Rozwiązanie zostanie przeprowadzone z wykorzystaniem siatki przedstawionej na rysunku 9.4. Problem jest dwuwymiarowy. Równanie opisujące analizowany problem to równanie Laplace'a w postaci

$$\lambda \nabla^2 T = 0, \quad \text{co oznacza} \quad \lambda \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) = 0$$

gdzie λ jest współczynnikiem przewodności cieplnej. Do dyskretyzacji równania Laplace'a zostanie użyta zależność (9.4) dająca przybliżenie rzędu $\mathcal{O}(h^2)$. Po wprowadzeniu siatki przedstawionej na rysunku 9.4 i użyciu odpowiednich indeksów, przybliżenia drugich pochodnych będą miały postać

$$\left.\frac{\partial^2 T}{\partial x^2}\right|_i = \frac{T_{i+1,j} - 2\,T_{i,j} + T_{i-1,j}}{h_x^2}, \qquad \left.\frac{\partial^2 T}{\partial y^2}\right|_i = \frac{T_{i,j+1} - 2\,T_{i,j} + T_{i,j-1}}{h_y^2}$$

Po użyciu takiej dyskretyzacji rozwiązywane równanie różniczkowe przyjmie postać

$$\lambda \left(\frac{T_{i+1,j} - 2\,T_{i,j} + T_{i-1,j}}{h_x^2} + \frac{T_{i,j+1} - 2\,T_{i,j} + T_{i,j-1}}{h_y^2} \right) = 0$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\frac{\lambda}{h_x^2}T_{i+1,j} - 2\lambda T_{i,j}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) + \frac{\lambda}{h_x^2}T_{i-1,j} + \frac{\lambda}{h_y^2}T_{i,j+1} + \frac{\lambda}{h_y^2}T_{i,j-1} = 0$$

Ponieważ przyjęto $h_x = h_y = h$, można zapisać równanie różnicowe

$$\frac{\lambda}{h^2}\left(T_{i+1,j} - 4T_{i,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}\right) = 0$$

Taka zależność różnicowa pozwoli przybliżyć wartości pochodnych we wnętrzu całego obszaru, to znaczy, dla węzłów, które na rysunku 9.4 mają jaśniejsze tło. Wektor niewiadomych ma postać

$$\mathbb{T} = \begin{bmatrix} T_1 & T_2 & T_3 & \dots & T_k & \dots & T_{33} \end{bmatrix}$$

gdzie odpowiednie numery węzłów, na podstawie rysunku 9.4, można wyrazić za pomocą indeksów, na przykład $T_3 = T_{51}$. Można teraz zapisać równania dla poszczególnych węzłów wewnętrznych, np. równanie dla węzła numer 6, dotyczące temperatury $T_6 = T_{32}$, będzie miało postać

$$\frac{\lambda}{h^2}\left(T_{4,2} - 4T_{3,2} + T_{2,2} + T_{3,3} + T_{3,1}\right) = 0$$

co, używając numerów węzłów, można zapisać

$$\frac{\lambda}{h^2}\left(T_7 - 4T_6 + T_5 + T_{12} + T_1\right) = 0$$

Dla każdego wewnętrznego węzła można zapisać podobne równanie. Nieco więcej uwagi należy poświęcić węzłom leżącym na brzegu analizowanego obszaru. Na fragmencie brzegu jest zadana wartość temperatury, na pozostałych częściach zdefiniowana jest wartość strumienia ciepła wymienianego z otoczeniem. Warunki brzegowe Dirichleta (wartość temperatury) przyjęte są dla węzłów 1, 5 i 10 oraz od 28 do 33. Nałożenie takich warunków brzegowych sprowadza się do zapisania układu równań typu

$$T_i = 0, \quad \text{gdzie} \quad i = 1, 5, 10$$

oraz

$$T_i = 50, \quad \text{gdzie} \quad i = 28, 29, \dots, 33$$

Warunki Neumana, określające wartości pochodnej funkcji na brzegu dotyczą pozostałych węzłów i wymagają specjalnego traktowania, ponieważ równania je opisujące mają postać

$$\dot{q} = -\lambda\frac{\partial T}{\partial \mathbb{n}}$$

Zatem dla węzłów leżących na brzegu z warunkiem brzegowym na pochodną normalną temperatury konieczne jest zapisanie równania wyrażającego tę pochodną.

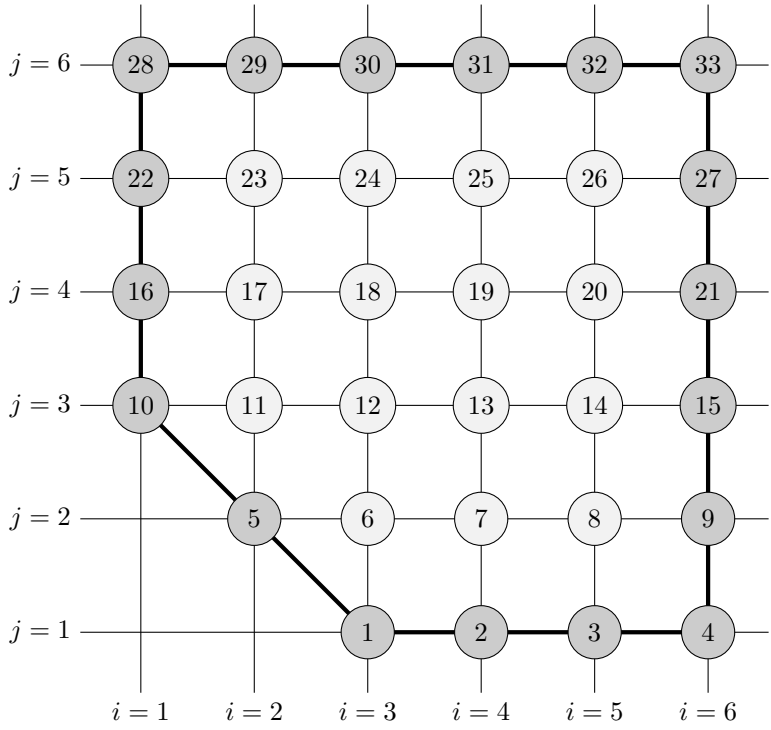
Można to zrobić na kilka sposobów:

— używając ilorazów różnicowych pierwszego rządu; ale mają one błąd rzędu $\mathcal{O}(h)$, co oznacza, że nastąpi degradacja jakości rozwiązania;

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— używając ilorazu różnicowego centralnego; ale w takiej sytuacji konieczne staje się dodanie fikcyjnego węzła poza obszarem analizy, by można było określić wartość tego ilorazu;
— wykorzystać wyrażenia trójpunktowe na pochodną na brzegu.

Ostatnia możliwość dotyczy wykorzystania wierszy numer 4 i 5 z tabeli 9.1. Na brzegu lewym i dolnym wykorzystane zostanie wyrażenie różnicowe z wiersza 4, na brzegu prawym wyrażenie z wiersza 5.



Rysunek 9.4. Siatka obliczeniowa metody różnic skończonych

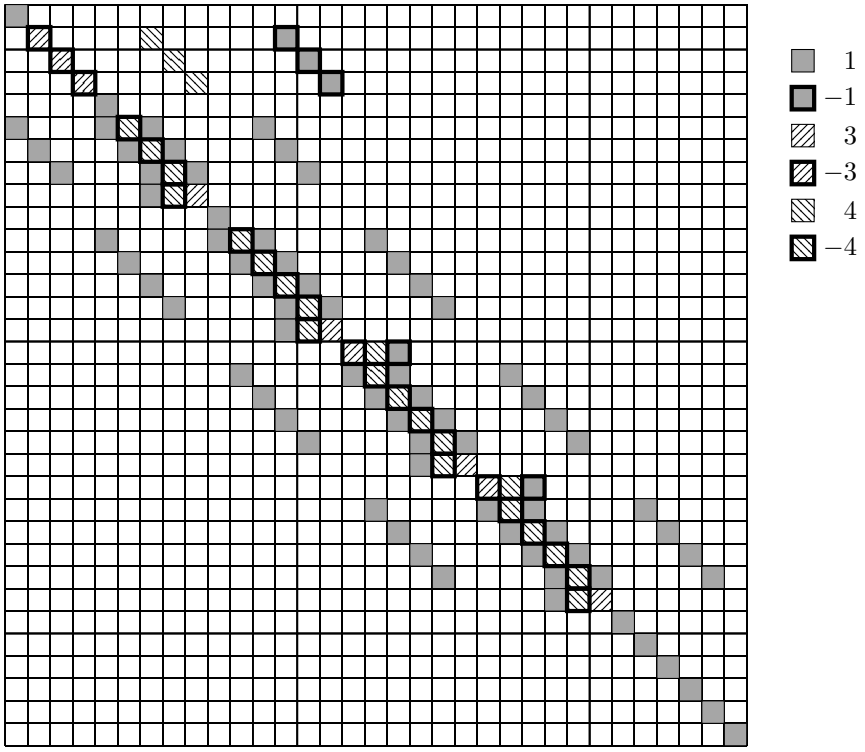
Dla przykładu zostaną zapisane równania dla węzłów numer 15 oraz 16, które będą stanowić odpowiednio 15 i 16 wiersz układu równań dla rozwiązywanego problemu.

$$\begin{aligned} T_{13} - 4T_{14} + 3T_{15} &= 0 \\ -3T_{16} + 4T_{17} - T_{18} &= 0 \end{aligned}$$

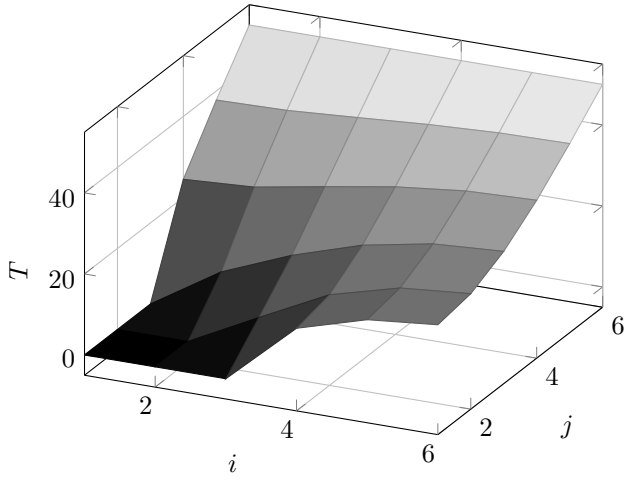
Z uwagi na znaczny rozmiar macierzy współczynników $\mathbb{A}$ została ona przedstawiona symbolicznie na rysunku 9.5. Każda niezerowa pozycja w macierzy została wypełniona kolorem lub kreskowaniem. Ujemne wartości współczynników zostały obwiedzione grubszą linią.

Wektor prawych stron układu $\mathbb{b}$ zawiera zera z wyjątkiem ostatnich 6 elementów, które mają wartość 50. Rozwiązanie problemu przedstawiono na rysunku 9.6.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 9.5. Macierz współczynników $\mathbb{A}$



Rysunek 9.6. Rozkład temperatur wyznaczony za pomocą MRS

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozwiązaniem układu równań jest wektor $\mathbb{T}$ zawierający 33 elementy stanowiące wartości temperatur w węzłach. Ponieważ sam wektor temperatur jest zbyt duży, by go przedstawić tu w postaci numerycznej, przedstawiono go graficznie na rysunku 9.6. Wyraźnie widać, że warunki brzegowe dotyczące temperatur zostały zachowane. Zaczerniony obszar o temperaturze 0 na lewo od linii wyznaczonej przez węzły 1, 5 i 10 wynika z procedury wizualizacyjnej. W tym obszarze wartości temperatur nie były wyznaczane.

9.6. Błędy

Rozwiązanie równania różniczkowego cząstkowego za pomocą metody różnic skończonych zazwyczaj jest przybliżone. Wynika to z występowania błędów zaburzających dokładność rozwiązania. Największy wpływ mają błędy zaokrągleń i błędy obcięcia. Obydwa rodzaje błędów były już omawiane w rozdziale wstępnym podręcznika.

Warto jedynie wspomnieć, że błąd obcięcia określa różnicę między dokładnym rozwiązaniem równania różniczkowego i jego aproksymacją za pomocą schematów różnic skończonych.

Można też mówić o błędzie dyskretyzacji. W metodzie różnic skończonych błąd dyskretyzacji jest bezpośrednio związany z błędem obcięcia.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 10

Metoda elementów brzegowych

10.1. Wstęp

Podstawową cechą różniącą metodę elementów brzegowych od metody elementów skończonych czy metody różnic skończonych jest sposób dyskretyzacji analizowanego obszaru. Zarówno w metodzie elementów skończonych, jak i różnic skończonych dyskretyzuje się cały obszar rozwiązania problemu. W metodzie elementów brzegowych dyskretyzuje się tylko brzeg. Rozwiązanie problemu otrzymuje się w zdefiniowanych punktach wewnątrz obszaru. Rozwiązanie w punktach jest dokładne – w odróżnieniu od pozostałych dwóch metod. Zatem głównymi korzyściami są:

— zmniejszenie wymiaru problemu,
— duża dokładność rozwiązania w punktach.

Pomimo tych niewątpliwych korzyści należy stwierdzić, iż nie zawsze można zastosować metodę elementów brzegowych. W niektórych wersjach, aby ją zastosować, musi być znane tzw. rozwiązanie fundamentalne równania określającego problem. Pod pojęciem rozwiązania fundamentalnego rozumie się funkcję spełniającą rozwiązywane jednorodne równanie różniczkowe, czyli z zerową prawą stroną, w całej przestrzeni z wyjątkiem pojedynczego punktu zwanego źródłem lub miejscem przyłożenia obciążenia. W źródle prawa strona równania opisującego zjawisko staje się nieskończona. Na przykład dla równania Laplace'a (10.1), opisującego rozkład temperatury rozwiązanie fundamentalne stanowi rozkład temperatury w materiale o jednostkowej przewodności cieplnej przy punktowym, stacjonarnym źródle ciepła.

$$\nabla^2 u = 0 \tag{10.1}$$

Są też wersje metody elementów brzegowych niewymagające znajomości rozwiązania fundamentalnego, wymienić można metodę Trefftza pozwalającą na rozwiązywanie problemów bez znajomości rozwiązania fundamentalnego [7].

Zasadniczo wyróżnia się dwa rodzaje metody elementów brzegowych: bezpośrednia oraz pośrednia. Prekursorem metody elementów brzegowych była metoda Trefftza, zaproponowana w 1926 roku do rozwiązywania problemu skręcania Saint-Venanta, polegająca na aproksymacji rozwiązania za pomocą superpozycji rozwiązań fundamentalnych równania różniczkowego opisującego problem [3].

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

10.1.1. Metoda Trefftza

W odróżnieniu od metody Rayleigha–Ritza oraz metody Galerkina (podrozdziały 11.4.1 oraz 11.4.2) w metodzie Trefftza funkcja aproksymująca nie spełnia warunków brzegowych, chociaż spełnia równanie różniczkowe. Nieznane parametry funkcji aproksymujących wyznacza się metodą residuów ważonych (reszt ważonych), czyli z warunku ortogonalności w odniesieniu do brzegu. Metoda Trefftza jest wykorzystywana do problemów z warunkami brzegowymi Dirichleta, dla których daje oszacowanie górne rozwiązania ścisłego.

Niech w obszarze Ω będzie zdefiniowany za pomocą równania różniczkowego problem

$$Lu - f = 0 \tag{10.2}$$

z warunkami brzegowymi na brzegu Γ

$$Bu - \hat{u} = 0 \tag{10.3}$$

Funkcje f oraz $\hat{u}$ są dane, natomiast L i B są operatorami różniczkowymi na poszukiwanej funkcji u. Rozwiązanie, które jest brane pod uwagę, ma postać sumy całki szczególnej φ_0 równania (10.2) oraz całki ogólnej równania jednorodnego $Lu = 0$

$$u_m = \sum_{i=1}^{m} c_i \varphi_i + \varphi_0 \tag{10.4}$$

Aby uzyskać pełną postać rozwiązania, należy wyznaczyć niewiadome parametry c_i. Ponieważ funkcje aproksymujące spełniają wyjściowe równanie różniczkowe, pozostaje kwestia spełnienia warunków brzegowych. Zatem wykorzystując metodę ważonych residuów dla warunków brzegowych (10.3), można napisać

$$\int_{\Gamma} (Bu_m - \hat{u}) \frac{\partial \varphi_i}{\partial \mathbb{n}} \mathrm{d}\Gamma = 0, \text{dla } i = 1, 2, \ldots, m \tag{10.5}$$

oczywiście $\frac{\partial \varphi_i}{\partial \mathbb{n}}$ jest pochodną normalną do brzegu z funkcji φ_i. Po podstawieniu poszukiwanego rozwiązania (10.4) do powyższej zależności otrzyma się następujące zależności

$$\int_{\Gamma} \left(\sum_{j=1}^{m} Bc_j \varphi_j - \hat{u} \right) \frac{\partial \varphi_i}{\partial \mathbb{n}} \mathrm{d}\Gamma, \text{dla } i = 1, 2, \ldots, m$$

$$\sum_{j=1}^{m} \left[c_j \int_{\Gamma} \frac{\partial \varphi_i}{\partial \mathbb{n}} B\varphi_j \mathrm{d}\Gamma \right] - \int_{\Gamma} \hat{u} \frac{\partial \varphi_i}{\partial \mathbb{n}} \mathrm{d}\Gamma = 0$$

czyli

$$\sum_{j=1}^{m} \left[c_j \int_{\Gamma} \frac{\partial \varphi_i}{\partial \mathbb{n}} B\varphi_j \mathrm{d}\Gamma \right] = \int_{\Gamma} \hat{u} \frac{\partial \varphi_i}{\partial \mathbb{n}} \mathrm{d}\Gamma \tag{10.6}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Powyższa zależność w rzeczywistości stanowi układ m równań z m niewiadomymi, które można jednoznacznie wyznaczyć.

Problem rozwiązania równania różniczkowego (10.2) został sprowadzony do wyznaczenia całek po brzegu obszaru, co stanowi sedno metody elementów brzegowych.

10.2. Podstawy matematyczne

Aby zrozumieć metodę elementów brzegowych, niezbędne jest przypomnienie pewnych podstawowych zależności matematycznych. Opisywane zależności dotyczą jednospójnego obszaru Ω z gładkim brzegiem $\partial\Omega$. Dla uproszczenia dalsze rozważania będą dotyczyły przypadków płaskich, które bez większych trudności można uogólnić.

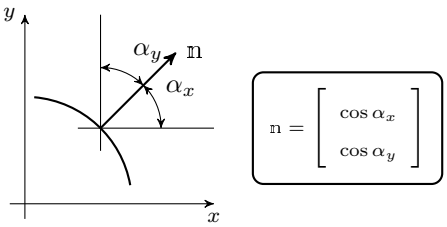
Twierdzenie Gaussa o dywergencji jest jednym z podstawowych twierdzeń związanych z analizą układów o parametrach rozłożonych. Nie będzie tu przytaczane w całości. Dywergencją wektora $\mathbb{f}$ (niektórzy nazywają ją rozbieżnością) nazywa się iloczyn skalarny $\nabla \cdot \mathbb{f}$.

$$\int_{\Omega} \nabla \cdot \mathbb{f} \mathrm{d}\Omega = \int_{\partial\Omega} \mathbb{f} \cdot \mathbb{n} \mathrm{d}s \tag{10.7}$$

Zależność (10.7) stanowi odpowiednik wzoru dotyczącego całkowania przez części.

Druga tożsamość Greena wynika bezpośrednio z zastosowania wzoru (10.7) do funkcji $\mathbb{f}$ w postaci $\mathbb{f} = \mathbb{g}\nabla\mathbb{h}$ oraz $\mathbb{f} = \mathbb{h}\nabla\mathbb{g}$. Uzyskane wyrażenia należy odjąć stronami. Sama tożsamość ma postać zapisaną zależnością (10.8).

$$\int_{\Omega} g\nabla^2 h - h\nabla^2 g \mathrm{d}\Omega = \int_{\partial\Omega} g\nabla h \cdot \mathbb{n} - h\nabla g \cdot \mathbb{n} \, \mathrm{d}\partial\Omega \tag{10.8}$$



Rysunek 10.1. Składowe wektora normalnego do brzegu

Dla przypomnienia, współrzędnymi wektora normalnego do brzegu $\mathbb{n}$ są kosinusy kierunkowe tego wektora. W przypadku płaskim, przedstawionym na rysunku 10.1, gdy wektor $\mathbb{n}$ tworzy z osiami układu współrzędnych kąty α_x oraz α_y, ma on postać $[\cos\alpha_x, \cos\alpha_y]^T$.

Zależności (10.7) oraz (10.8) można też stosować do obszarów wielospójnych i brzegów odcinkami gładkich.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

10.3. Równania całkowe metody elementów brzegowych

Metoda elementów brzegowych prowadzi do dyskretyzacji brzegu. Podobnie jak w metodzie Trefftza wykorzystywane są funkcje spełniające równanie różniczkowe. Te funkcje są zwane rozwiązaniami fundamentalnymi. Chociaż używane są ścisłe rozwiązania równania różniczkowego, to jednak metoda daje wyniki przybliżone, ponieważ warunki brzegowe są aproksymowane. Niemniej, jej zaletą jest to, że dyskretyzacja odbywa się jedynie na brzegu. W zagadnieniach, w których występują źródła, trzeba też wyznaczyć całkę po obszarze rozwiązania, ale dyskretyzacja nie musi być aż tak dokładna jak w przypadku metody elementów skończonych.

Podstawowe pojęcia związane z metodą elementów brzegowych zostaną przedstawione na przykładzie rozwiązania równania różniczkowego cząstkowego opisującego ustalony w czasie przepływ ciepła.

10.3.1. Ustalony przepływ ciepła

Ustalony przepływ ciepła w izotropowym obszarze Ω, ze współczynnikiem przewodnictwa cieplnego λ będącym funkcją temperatury i położenia, opisuje równanie (10.9) [32].

$$\frac{\partial}{\partial x}\left(\lambda\frac{\partial T}{\partial x}\right)+\frac{\partial}{\partial y}\left(\lambda\frac{\partial T}{\partial y}\right)+\frac{\partial}{\partial z}\left(\lambda\frac{\partial T}{\partial z}\right)+Q(x,y,z)=0 \tag{10.9}$$

Gdy współczynnik λ jest stały, równanie przyjmuje postać

$$\lambda\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}\right)+Q(x,y,z)=0 \tag{10.10}$$

Zmienna T oznacza temperaturę w kelwinach lub stopniach Celsjusza, λ jest współczynnikiem przewodzenia ciepła w $[\mathrm{W{\cdot}m^{-1}{\cdot}K^{-1}}]$, zaś Q jest źródłem ciepła wyrażonym dla przypadku płaskiego w $[\mathrm{W{\cdot}m^{-2}}]$.

Aby było możliwe rozwiązanie tego równania, konieczne jest nałożenie warunków brzegowych. Dla zagadnień przepływu ciepła najczęściej przyjmuje się następujące warunki brzegowe:

— Dirichleta: $T(x)=T_b$, gdzie T_b jest znaną temperaturą,
— Neumana: $q(x)=-\lambda\frac{\partial T}{\partial \mathtt{n}}=q_b$, gdzie q_b jest strumieniem ciepła o znanej gęstości,
— Robina: $q(x)=-\lambda\frac{\partial T}{\partial \mathtt{n}}=\alpha(T-T_{ot})$, gdzie T_{ot} jest znaną temperaturą otoczenia, natomiast α jest współczynnikiem wnikania ciepła.

Pochodną normalną do brzegu należy rozumieć w następujący sposób

$$\frac{\partial T}{\partial \mathtt{n}}=\frac{\partial T}{\partial x}\cos\alpha_x+\frac{\partial T}{\partial y}\cos\alpha_y=\nabla T\cdot\mathtt{n} \tag{10.11}$$

gdzie, jak pokazano na rysunku 10.1, α_x oraz α_y są kątami między kierunkiem normalnej do brzegu, oznaczanej przez $\mathtt{n}$ a osiami układu współrzędnych, odpowiednio x i y.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wychodząc z metody residuów ważonych, można oczekiwać, że rozwiązanie równania (10.10) będzie spełniać warunek

$$\int_{\Omega} \left(\lambda \nabla^2 T(x) + Q(x)\right) T^{\star}(\xi, x) \mathrm{d}\Omega = 0 \tag{10.12}$$

gdzie funkcja wagowa $T^{\star}(\xi, x)$ stanowi tak zwane rozwiązanie fundamentalne, $\xi = (\xi_x, \xi_y)$ jest punktem zwanym punktem obserwacji, natomiast $x = (x_x, x_y)$ jest nazywany punktem całkowania. Rozwiązanie fundamentalne jest rozwiązaniem równania w postaci

$$\lambda \nabla^2 T^{\star}(\xi, x) = -\delta(\xi, x) \tag{10.13}$$

gdzie $\delta(\xi, x)$ jest deltą Diraca definiowaną następująco

$$\delta(\xi, x) = \begin{cases} 0, \text{ gdy } (\xi_x, \xi_y) \neq (x_x, x_y) \\ \infty, \text{ gdy } (\xi_x, \xi_y) = (x_x, x_y) \end{cases}$$

Ważne jest, że

$$\int_{\Omega} \delta(\xi, x) \mathrm{d}\Omega = 1$$

z czego z kolei wynika

$$\int_{\Omega} T(x) \delta(\xi, x) \mathrm{d}\Omega = T(\xi) \tag{10.14}$$

Rozwiązanie fundamentalne równania (10.10) w przypadku dwuwymiarowym stanowi funkcja

$$T^{\star}(\xi, x) = \frac{1}{2\pi\lambda} \ln \frac{1}{r} \tag{10.15}$$

gdzie r oznacza odległość między punktem obserwacji ξ a punktem całkowania x. Zostanie wprowadzona nowa funkcja $q^{\star}(\xi, x)$ stanowiąca pochodną normalną rozwiązania fundamentalnego

$$\begin{aligned} q^{\star}(\xi, x) = -\lambda \frac{\partial T^{\star}(\xi, x)}{\partial \mathrm{n}} = -\lambda \left(\frac{\partial T^{\star}(\xi, x)}{\partial x} \cos \alpha_x + \frac{\partial T^{\star}(\xi, x)}{\partial y} \cos \alpha_y \right) = \\ = \frac{(x_x - \xi_x) \cos \alpha_x + (x_y - \xi_y) \cos \alpha_y}{2\pi r^2} \end{aligned} \tag{10.16}$$

Po tym krótkim wyjaśnieniu można zapisać prawą stronę równania (10.12) w następującej postaci

$$\int_{\Omega} \left(\lambda \nabla^2 T(x) + Q\right) T^{\star}(\xi, x) \mathrm{d}\Omega = \int_{\Omega} \lambda \nabla^2 T(x) T^{\star}(\xi, x) \mathrm{d}\Omega + \int_{\Omega} Q(x) T^{\star}(\xi, x) \mathrm{d}\Omega \tag{10.17}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Do pierwszej całki po prawej stronie można zastosować drugą tożsamość Greena (10.8), co można zapisać jako

$$\int_{\Omega} \lambda T^{\star}(\xi,x)\nabla^2 T(x)\mathrm{d}\Omega - \int_{\Omega} \lambda T(x)\nabla^2 T^{\star}(\xi,x)\mathrm{d}\Omega =$$
$$= \int_{\partial\Omega} \lambda T^{\star}(\xi,x)\nabla T(x)\cdot \mathtt{n} - \lambda T(x)\nabla T^{\star}(\xi,x)\cdot \mathtt{n}\mathrm{d}\partial\Omega =$$
$$= \int_{\partial\Omega} T^{\star}(\xi,x)\underbrace{\lambda\frac{\partial T(x)}{\partial \mathtt{n}}}_{-q(x)} - T(x)\underbrace{\lambda\frac{\partial T^{\star}(\xi,x)}{\partial \mathtt{n}}}_{-q^{\star}(\xi,x)}\mathrm{d}\partial\Omega \quad (10.18)$$

Co po uporządkowaniu daje

$$\int_{\Omega} \lambda T^{\star}(\xi,x)\nabla^2 T(x)\mathrm{d}\Omega = \int_{\Omega} \lambda T(x)\nabla^2 T^{\star}(\xi,x)\mathrm{d}\Omega +$$
$$+ \int_{\partial\Omega} -T^{\star}(\xi,x)q(x) + T(x)q^{\star}(\xi,x)\mathrm{d}\partial\Omega \quad (10.19)$$

Zatem równanie (10.17) można zapisać w postaci

$$\int_{\Omega} \left(\lambda\nabla^2 T(x) + Q(x)\right) T^{\star}(\xi,x)\mathrm{d}\Omega = \int_{\Omega} \lambda T(x)\nabla^2 T^{\star}(\xi,x)\mathrm{d}\Omega - \int_{\partial\Omega} T^{\star}(\xi,x)q(x)\mathrm{d}\partial\Omega +$$
$$+ \int_{\partial\Omega} T(x)q^{\star}(\xi,x)\mathrm{d}\partial\Omega + \int_{\Omega} Q(x)T^{\star}(\xi,x)\mathrm{d}\Omega \quad (10.20)$$

Wobec tego równanie (10.12), stanowiące całkowy odpowiednik równania (10.10), można zapisać w postaci

$$\underbrace{\int_{\Omega} T(x)\underbrace{\lambda\nabla^2 T^{\star}(\xi,x)}_{-\delta(\xi,x)}\mathrm{d}\Omega}_{-T(\xi)} - \int_{\partial\Omega} T^{\star}(\xi,x)q(x)\mathrm{d}\partial\Omega +$$
$$+ \int_{\partial\Omega} T(x)q^{\star}(\xi,x)\mathrm{d}\partial\Omega + \int_{\Omega} Q(x)T^{\star}(\xi,x)\mathrm{d}\Omega = 0 \quad (10.21)$$

Po uwzględnieniu równań (10.13) oraz (10.14) można zapisać postać równania całkowego dla każdego punktu wewnątrz obszaru

$$T(\xi) + \int_{\partial\Omega} T^{\star}(\xi,x)q(x)\mathrm{d}\partial\Omega = \int_{\partial\Omega} T(x)q^{\star}(\xi,x)\mathrm{d}\partial\Omega + \int_{\Omega} Q(x)T^{\star}(\xi,x)\mathrm{d}\Omega \quad (10.22)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Gdy punkt obserwacji leży na brzegu, równanie brzegowe przyjmuje postać

$$B(\xi)T(\xi) + \int_{\partial\Omega} T^{\star}(\xi, x)q(x)\mathrm{d}\partial\Omega = \int_{\partial\Omega} T(x)q^{\star}(\xi, x)\mathrm{d}\partial\Omega + \int_{\Omega} Q(x)T^{\star}(\xi, x)\mathrm{d}\Omega \quad (10.23)$$

gdzie $B(\xi)$ jest kątem wewnętrznym obszaru w punkcie ξ. Można go traktować jak współczynnik zależny od kształtu brzegu. Na gładkim fragmencie brzegu $B(\xi) = \frac{1}{2}$. Oczywiście, jeżeli nie ma źródeł ciepła wewnątrz obszaru, to ostatnie całki w równaniach (10.22) oraz (10.23) się zerują i równania brzegowe przyjmują postać, w której nie występują całkowania po obszarze.

10.3.2. Numeryczna aproksymacja równań brzegowych

Po założeniu, że w analizowanym obszarze nie występują źródła ciepła, równania całkowe (10.22) oraz (10.23) odpowiednio przyjmują postać (10.24) i (10.25).

$$T(\xi) + \int_{\partial\Omega} T^{\star}(\xi, x)q(x)\mathrm{d}\partial\Omega = \int_{\partial\Omega} T(x)q^{\star}(\xi, x)\mathrm{d}\partial\Omega \quad (10.24)$$

$$B(\xi)T(\xi) + \int_{\partial\Omega} T^{\star}(\xi, x)q(x)\mathrm{d}\partial\Omega = \int_{\partial\Omega} T(x)q^{\star}(\xi, x)\mathrm{d}\partial\Omega \quad (10.25)$$

Rozwiązując problem metodą elementów brzegowych, przede wszystkim należy zdecydować o sposobie dyskretyzacji brzegu. Dyskretyzacja brzegu obszaru oznacza podział brzegu na elementy, wobec tego

$$\partial\Omega = \sum_{j=1}^{N} \partial\Omega_j$$

Taka dyskretyzacja powoduje, że całkowe równania brzegowe przechodzą w układ liniowych równań algebraicznych. A zatem równanie (10.25) dla ustalonego punktu ξ_i przyjmuje postać

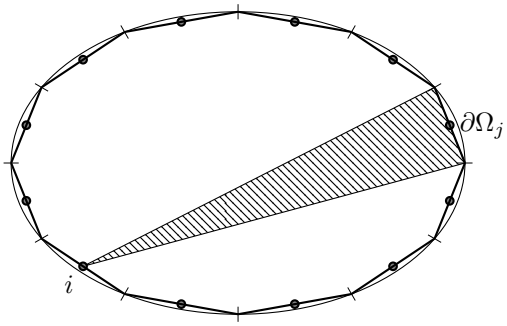
$$B(\xi_i)T(\xi_i) + \sum_{j=1}^{N} \int_{\partial\Omega_j} T^{\star}(\xi_i, x)q(x)\mathrm{d}\partial\Omega_j = \sum_{j=1}^{N} \int_{\partial\Omega_j} T(x)q^{\star}(\xi_i, x)\mathrm{d}\partial\Omega_j \quad (10.26)$$

Podobnie przekształca się równanie (10.24) dotyczące punktów ξ leżących wewnątrz obszaru Ω. Aby rozwiązać uzyskane równania brzegowe, należy przyjąć sposób aproksymacji warunków brzegowych dla poszczególnych elementów. Zazwyczaj wykorzystuje się elementy stałe, liniowe lub paraboliczne. Elementy stałe, przedstawione na rysunku 10.2 mają jeden węzeł w środku, elementu a wartości temperatury i strumienia ciepła na długości całego elementu są stałe.

Gdy przyjmie się dyskretyzację elementami stałymi, równanie (10.26) można zapisać w prostszej postaci jako

$$B(\xi_i)T(\xi_i) + \sum_{j=1}^{N} q_j \int_{\partial\Omega_j} T^{\star}(\xi_i, x)\mathrm{d}\partial\Omega_j = \sum_{j=1}^{N} T_j \int_{\partial\Omega_j} q^{\star}(\xi_i, x)\mathrm{d}\partial\Omega_j \quad (10.27)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 10.2. Elementy stałe

Do wyznaczenia pozostają całki z rozwiązania fundamentalnego i pochodnej normalnej tego rozwiązania. Niech ich wartości będą oznaczone jako

$$g_{ij} = \int\limits_{\partial\omega_j} T^\star(\xi_i, x)\mathrm{d}\partial\Omega_j, \quad \text{oraz} \quad h_{ij} = \int\limits_{\partial\omega_j} q^\star(\xi_i, x)\mathrm{d}\partial\Omega_j \tag{10.28}$$

Wartości g_{ij} oraz h_{ij} stanowią elementy macierzy wpływu. Najczęściej wyznacza się je numerycznie.

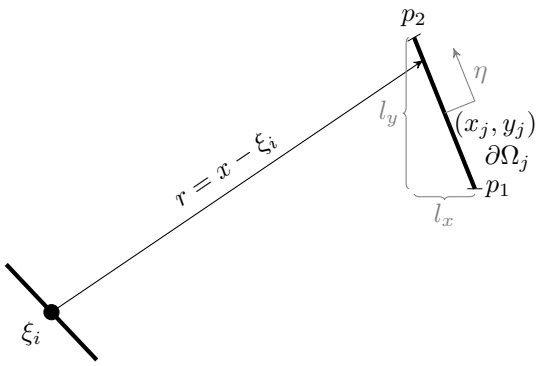
Elementy macierzy wpływu. Aby wyznaczyć poszczególne elementy macierzy wpływu na podstawie zależności (10.28) można wykorzystać kwadratury Gaussa. Wobec tego konieczna jest zamiana zmiennych na współrzędną $\eta \in \langle -1, 1\rangle$. Punkt początkowy odcinka zostanie oznaczony przez p_1, zaś końcowy przez p_2. Współrzędne środka odcinka $\partial\Omega_j$ są oznaczone przez (x_j, y_j). Zatem współrzędne w układzie globalnym wyrażają się przez współrzędne lokalne w sposób następujący

$$\begin{cases} x_x = \frac{1-\eta}{2}p_{1x} + \frac{1+\eta}{2}p_{2x} = x_j + \frac{l_x}{\eta} \\ x_y = \frac{1-\eta}{2}p_{1y} + \frac{1+\eta}{2}p_{2y} = y_j + \frac{l_y}{\eta} \end{cases} \tag{10.29}$$

Po zamianie zmiennych $\mathrm{d}\partial\Omega_j = \frac{l}{2}\mathrm{d}\eta$ poszczególne całki będą miały postacie przedstawione w równaniach (10.30) oraz (10.31).

$$\begin{aligned} g_{ij} = \int\limits_{\partial\omega_j} T^\star(\xi_i, x)\mathrm{d}\partial\Omega_j &= \int\limits_{\partial\omega_j} \frac{1}{2\pi\lambda}\ln\frac{1}{r}\mathrm{d}\partial\Omega_j = \\ &= \frac{1}{2\pi\lambda}\int\limits_{\partial\Omega_j} \ln\frac{1}{\sqrt{(x_x-\xi_{ix})^2+(x_y-\xi_{iy})^2}}\mathrm{d}\partial\Omega_j = \\ &= \frac{l}{4\pi\lambda}\int\limits_{-1}^{1} \ln\frac{1}{\sqrt{(x_j+\frac{l_x}{2}\eta-\xi_{ix})^2+(y_j+\frac{l_y}{2}\eta-\xi_{iy})^2}}\mathrm{d}\eta \end{aligned} \tag{10.30}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 10.3. Zależności geometryczne przy wyliczaniu wartości g_{ij} oraz h_{ij}

Po porównaniu rysunków 10.1 oraz 10.3, składowe wektora normalnego $\mathbb{n}$ można wyrazić zależnością

$$\mathbb{n} = \left[\begin{array}{cc} \frac{p_{2y}-p_{1y}}{\sqrt{(p_{2x}-p_{1x})^2+(p_{2y}-p_{1y})^2}} & \frac{p_{2x}-p_{1x}}{\sqrt{(p_{2x}-p_{1x})^2+(p_{2y}-p_{1y})^2}} \end{array} \right]^T = \left[\begin{array}{cc} \frac{l_y}{l} & \frac{l_x}{l} \end{array} \right]^T$$

Zatem całka h_{ij} wyrazi się wzorem

$$\begin{aligned} h_{ij} = \int\limits_{\partial\omega_j} q^\star(\xi_i, x)\mathrm{d}\partial\Omega_j &= \int\limits_{\partial\omega_j} \frac{(x_x - \xi_{ix})\cos\alpha_x + (x_y - \xi_{iy})\cos\alpha_y}{2\pi r^2}\mathrm{d}\partial\Omega_j = \\ &= \frac{1}{2\pi}\int\limits_{\partial\omega_j} \frac{(x_x - \xi_{ix})\cos\alpha_x + (x_y - \xi_{iy})\cos\alpha_y}{(x_x - \xi_{ix})^2 + (x_y - \xi_{iy})^2}\mathrm{d}\partial\Omega_j = \\ &= \frac{1}{4\pi}\int\limits_{-1}^{1} \frac{(x_j + \frac{l_x}{2}\eta - \xi_{ix})l_y + (x_j + \frac{l_y}{2} - \xi_{iy})l_x}{(x_j + \frac{l_x}{2}\eta - \xi_{ix})^2 + (y_j + \frac{l_y}{2}\eta - \xi_{iy})^2}\mathrm{d}\eta \end{aligned} \quad (10.31)$$

Dla elementów, w których wartości temperatury i strumienia ciepła są stałe na całej długości elementu, wartości g_{ij} oraz h_{ij} można wyznaczyć analitycznie. Dla innych rodzajów elementów konieczne staje się wykorzystanie metod numerycznych.

W metodzie elementów brzegowych poważne problemy pojawiają się, gdy punkt obserwacji ξ_j pokrywa się z punktem całkowania x. W takiej sytuacji odległość r we wzorach (10.15) oraz (10.16) zdąża do zera, a całki wyrażające g_{ij} oraz h_{ij} stają się osobliwe. W ogólnym przypadku rozwiązanie tego problemu wymaga specjalnego traktowania tych całek. Jednak w analizowanym przypadku można obie te całki wyznaczyć analitycznie. Dla $i = j$ wartości omawianych całek przedstawiają się następująco

$$g_{ii} = \frac{l}{2\pi\lambda}\left(1 + \ln\frac{2}{l}\right), \quad \text{oraz} \quad h_{ii} = 0 \quad (10.32)$$

Należy też pamiętać, że współczynnik $B_i = B(\xi_i)$ jest kątem wewnętrznym obszaru w punkcie ξ_i i dla brzegów gładkich przyjmuje wartość 0,5.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Macierze wpływu i budowa układu równań. Równanie (10.26) przedstawia układ liniowych równań algebraicznych. Układ ten można zapisać w postaci

$$\mathbb{G}\mathbb{q} = \mathbb{H}\mathbb{T} - \delta_{ij}B_i \tag{10.33}$$

Elementami macierzy $\mathbb{G}$ są wartości g_{ij}, zaś elementami macierzy $\mathbb{H}$ są wcześniej wyznaczone wartości h_{ij}. Wektory $\mathbb{q}$ oraz $\mathbb{T}$ zawierają znane wartości wynikające z warunków brzegowych oraz niewiadome, które zostaną wyznaczone w trakcie rozwiązywania układu równań.

Ponieważ wektory $\mathbb{q}$ oraz $\mathbb{T}$ zawierają wartości znane i poszukiwane, wygodniej jest dokonać takiego przestawienia kolumn, by po lewej stronie układu równań znajdowały się jedynie te kolumny, które dotyczą nieznanych wartości w węzłach elementów brzegowych. Takie postępowanie prowadzi do układu równań

$$\mathbb{A}\mathbb{x} = \mathbb{B}\mathbb{y} \tag{10.34}$$

gdzie macierz $\mathbb{A}$ zawiera współczynniki przy poszukiwanych wartościach w węzłach, zarówno temperatury, jak i strumieni ciepła. Macierz $\mathbb{B}$ zawiera współczynniki stojące przy znanych wartościach, wynikających z warunków brzegowych, które zostały zapisane w wektorze $\mathbb{y}$. W ten sposób iloczyn po prawej stronie daje kolumnowy wektor znanych wartości. Niestety macierz $\mathbb{A}$ jest macierzą kwadratową, pełną i niesymetryczną, co powoduje, że jedyną efektywną metodą rozwiązywania układu równań (10.34) jest eliminacja Gaussa, zaś dla dużych układów równań metody iteracyjne.

Wyznaczenie wartości wewnątrz obszaru. Po wyznaczeniu wartości strumieni ciepła q_i i temperatury T_i w każdym węźle położonym na brzegu można wyznaczyć rozkład temperatury wewnątrz analizowanego obszaru. Do tego celu wykorzystuje się równanie (10.24). Podobnie jak poprzednio dyskretyzacja brzegu prowadzi do przekształcenia ciągłego równania całkowego w postać dyskretną. Tok przekształceń jest następujący

$$T(\xi_i) + \sum_{j=1}^{N} \int\limits_{\partial\Omega_j} T^\star(\xi_i, x) q(x) \mathrm{d}\partial\Omega_j = \sum_{j=1}^{N} \int\limits_{\partial\Omega_j} T(x) q^\star(\xi_i, x) \mathrm{d}\partial\Omega_j \tag{10.35}$$

Po przyjęciu elementów brzegowych stałych uzyskuje się postać

$$T(\xi_i) + \sum_{j=1}^{N} q_j \int\limits_{\partial\Omega_j} T^\star(\xi_i, x) \mathrm{d}\partial\Omega_j = \sum_{j=1}^{N} T_j \int\limits_{\partial\Omega_j} q^\star(\xi_i, x) \mathrm{d}\partial\Omega_j \tag{10.36}$$

Natomiast równanie służące do wyznaczania wartości temperatury w punktach wewnętrznych obszaru Ω ma postać (10.37).

$$T(\xi_i) = \sum_{j=1}^{N} T_j h_{ij} - \sum_{j=1}^{N} q_j g_{ij} \tag{10.37}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zatem by wyznaczyć wartości temperatury wewnątrz analizowanego obszaru, niezbędne jest realizowanie kolejnych całkowań po brzegu obszaru. Ponieważ współczynniki h_{ij} oraz g_{ij} są wyrażane całkami, w których funkcje podcałkowe zależą od odległości analizowanego punktu od poszczególnych elementów brzegu, to dla każdego punktu wewnątrz obszaru należy je wyznaczać na nowo.

Aby zilustrować te teoretyczne rozważania, zostanie przedstawiony praktyczny przykład analizy rozkładu temperatury.

10.4. Przykład obliczeniowy

Sformułowanie problemu. Wyznaczyć rozkład temperatury w stanie ustalonym w obszarze przedstawionym na rysunku 9.3. Wykorzystać metodę elementów brzegowych ze stałymi elementami brzegowymi. Zastosować dyskretyzację podobną do zastosowanej przy metodzie różnic skończonych.

Rozwiązanie zostanie przeprowadzone z wykorzystaniem siatki przedstawionej na rysunku 10.4. Problem jest dwuwymiarowy. Węzły brzegowe zaznaczono w postaci czarnych kropek, zaś węzły wewnętrzne zaznaczono na szaro. Elementy od 1 do 16 mają długość 0,1 m, zaś elementy 17 i 18 mają długość 1,4142 m. Równanie opisujące analizowany problem jest równaniem Laplace'a w postaci

$$\lambda \nabla^2 T = 0, \quad \text{co oznacza} \quad \lambda \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) = 0$$

gdzie λ jest współczynnikiem przewodności cieplnej. Zadane warunki brzegowe, przypisane kolejnym węzłom zestawiono w tabeli 10.1.

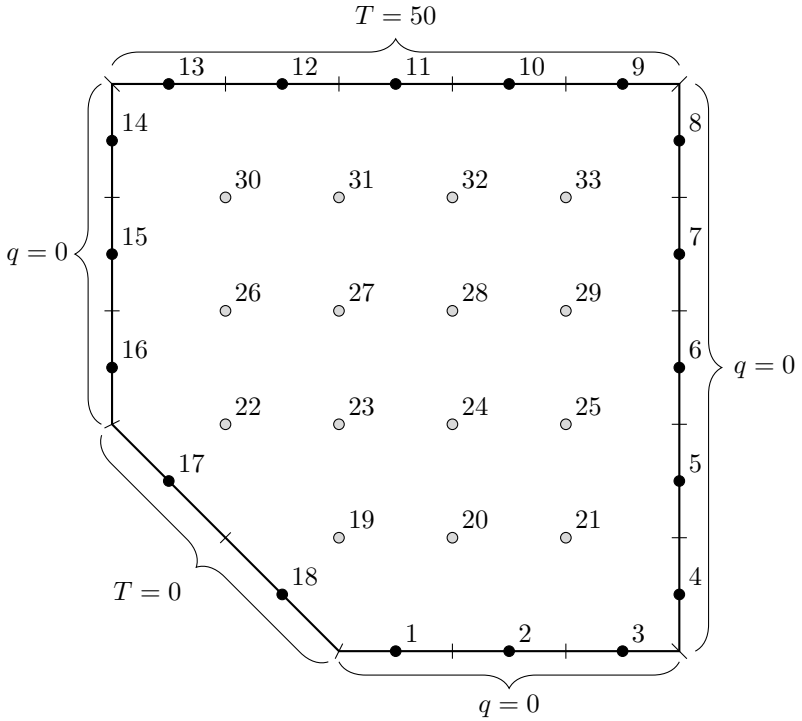
Tabela 10.1. Zadane wartości temperatury i strumieni ciepła w poszczególnych węzłach

Lp.	Węzły	Wartość	Jednostka
1.	od 9 do 13	50	°C
2.	17 i 18	0	°C
3.	od 1 do 8, od 14 do 15	0	W/m^2

Algorytm postępowania został przedstawiony na rysunku 10.5. Po prawej stronie zamieszczono informacje na temat tego, z czym wiąże się zrealizowane danego kroku w algorytmie.

W pierwszej kolejności należy wyznaczyć wartości temperatury i strumieni ciepła na brzegu. Aby tego dokonać, konieczne jest przypisanie węzłom znanych wartości czy to temperatury, czy strumieni ciepła.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 10.4. Siatka obliczeniowa metody elementów brzegowych

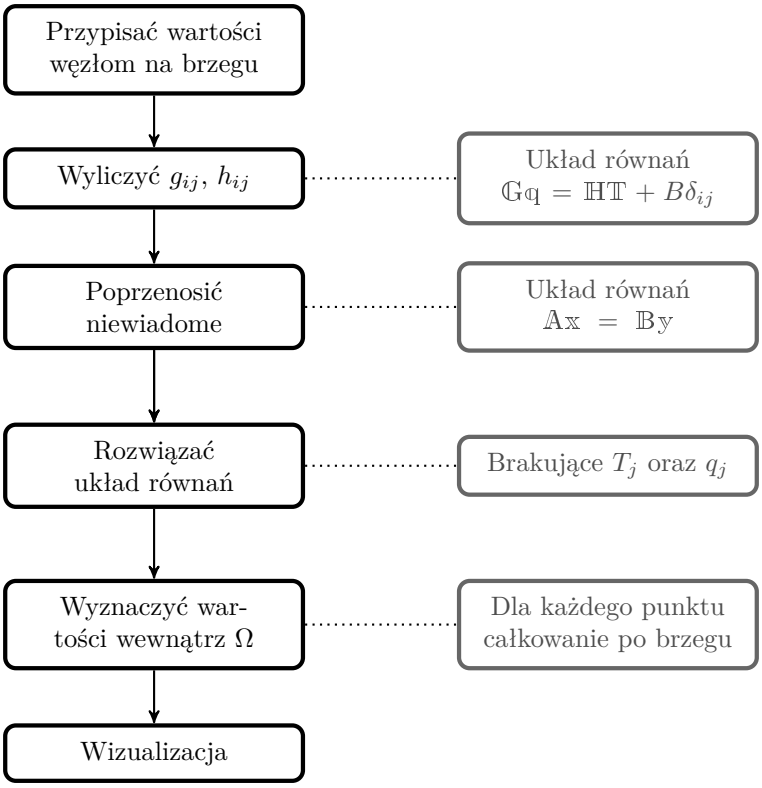
Następnie należy wyznaczyć elementy macierzy wpływu $\mathbb{G}$ oraz $\mathbb{H}$. Po wyznaczeniu potrzebnych wartości uzyskuje się układ równań w postaci (10.33). Ponieważ w macierzach $\mathbb{G}$ i $\mathbb{H}$ znajdują się kolumny dotyczące wartości znanych i poszukiwanych, konieczne jest uporządkowanie tych macierzy prowadzące do równania (10.34). Po rozwiązaniu uzyskuje się w każdym węźle brzegowym wartość temperatury oraz strumienia ciepła. Dzięki temu można wykorzystać zależność (10.37) do wyznaczenia wartości w węzłach wewnętrznych. Warunki brzegowe dla temperatury i strumienia ciepła zostały zapisane w zależnościach (10.38) oraz (10.39).

$$\mathbb{q} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & - & - & - & - & - & 0 & 0 & 0 & - & - \end{bmatrix}^T \tag{10.38}$$

$$\mathbb{T} = \begin{bmatrix} - & - & - & - & - & - & - & - & 50 & 50 & 50 & 50 & 50 & - & - & - & 0 & 0 \end{bmatrix}^T \tag{10.39}$$

Bez trudności można się zorientować, że w danym węźle na brzegu występuje warunek Dirichleta lub warunek Neumanna. Liczba znanych wartości brzegowych jest równa liczbie węzłów, co pozwala zbudować układ równań, który umożliwi wyznaczenie pozostałych wartości w węzłach.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356


Przypisać wartości węzłom na brzegu
Wyliczyć g_{ij}, h_{ij}
Układ równań $\mathbb{G}\mathbb{q} = \mathbb{H}\mathbb{T} + B\delta_{ij}$
Poprzenosić niewiadome
Układ równań $\mathbb{A}\mathbb{x} = \mathbb{B}\mathbb{y}$
Rozwiązać układ równań
Brakujące T_j oraz q_j
Wyznaczyć wartości wewnątrz Ω
Dla każdego punktu całkowanie po brzegu
Wizualizacja

Rysunek 10.5. Algorytm obliczeń metodą elementów brzegowych

Według algorytmu przedstawionego na rysunku 10.5 następnym krokiem jest wyznaczenie wartości elementów macierzy $\mathbb{H}$ oraz $\mathbb{G}$ zgodnie z zależnościami (10.30), (10.31) oraz (10.32).

Obliczenia wartości poszczególnych elementów macierzy wykonuje się wierszami. W i-tym wierszy znajdują się współczynniki wpływu dla i-tego elementu. Do całkowania można wykorzystać kwadratury Gaussa opisane w podrozdziale 4.7 w obliczeniach wykorzystano kwadraturę sześciopunktową. Realizując iteracje po kolejnych elementach, wylicza się kolejno współczynniki g_{ij} oraz h_{ij} z zależności (10.30), (10.31) oraz (10.32). Dla unaocznienia zostaną przeprowadzone obliczenia wartości pierwszego wiersza macierzy $\mathbb{G}$ oraz $\mathbb{H}$ dla elementu 8:

— pierwszy wiersz i ósmy element oznacza, że $i = 1$, $j = 8$,
— współrzędne węzła ξ_1 to (0,25; 0),
— współrzędne punktów $p_1 = (0{,}5;\ 0{,}4)$, $p_2 = (0{,}5;\ 0{,}5)$ zatem $l_x = 0$, $l_y = 0{,}1$,
— współrzędne węzłów Gaussa i wartości funkcji podcałkowych dla tych współrzędnych zestawiono w tabeli 10.2,
— wartości całek $g_{18} = 0{,}010581$ oraz $h_{18} = 0{,}015112$.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 10.2. Węzły Gaussa i wartości funkcji podcałkowych dla g_{18} oraz h_{18}

k	η_k	g_{ijk}	h_{ijk}	Wagi w_k
1.	-0,93247	0,74536	0,111006	0,17132
2.	-0,66121	0,72124	0,105780	0,36076
3.	-0,23862	0,68441	0,098269	0,46791
4.	0,23862	0,64389	0,090619	0,46791
5.	0,66121	0,60896	0,084503	0,36076
6.	0,93247	0,58699	0,080871	0,17132

$\mathbb{G} = 10^{-2}\cdot$

$$\begin{bmatrix}
6{,}36 & 3{,}74 & 2{,}58 & 2{,}17 & 1{,}96 & 1{,}65 & 1{,}34 & 1{,}06 & \overbrace{0{,}98 & 1{,}07 & 1{,}10 & 1{,}07 & 0{,}98 & 1{,}06}^{\text{Nieznane } q} & 1{,}34 & 1{,}65 & \overbrace{3{,}15 & 5{,}05}^{\text{Nieznane } q} \\
3{,}74 & 6{,}36 & 3{,}74 & 2{,}91 & 2{,}47 & 1{,}97 & 1{,}54 & 1{,}19 & 1{,}07 & 1{,}10 & 1{,}07 & 0{,}98 & 0{,}86 & 0{,}89 & 1{,}12 & 1{,}34 & 2{,}47 & 3{,}57 \\
2{,}58 & 3{,}74 & 6{,}36 & 4{,}20 & 2{,}96 & 2{,}18 & 1{,}66 & 1{,}26 & 1{,}10 & 1{,}07 & 0{,}98 & 0{,}86 & 0{,}71 & 0{,}72 & 0{,}89 & 1{,}06 & 1{,}92 & 2{,}69 \\
2{,}18 & 2{,}96 & 4{,}20 & 6{,}36 & 3{,}74 & 2{,}58 & 1{,}92 & 1{,}46 & 1{,}26 & 1{,}18 & 1{,}06 & 0{,}89 & 0{,}72 & 0{,}71 & 0{,}86 & 0{,}98 & 1{,}75 & 2{,}36 \\
1{,}97 & 2{,}47 & 2{,}91 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}58 & 1{,}92 & 1{,}65 & 1{,}53 & 1{,}34 & 1{,}12 & 0{,}89 & 0{,}86 & 0{,}98 & 1{,}07 & 1{,}80 & 2{,}27 \\
1{,}65 & 1{,}96 & 2{,}17 & 2{,}58 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}58 & 2{,}17 & 1{,}96 & 1{,}65 & 1{,}34 & 1{,}06 & 0{,}98 & 1{,}07 & 1{,}10 & 1{,}74 & 2{,}04 \\
1{,}34 & 1{,}53 & 1{,}65 & 1{,}92 & 2{,}58 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}91 & 2{,}47 & 1{,}97 & 1{,}54 & 1{,}19 & 1{,}07 & 1{,}10 & 1{,}07 & 1{,}59 & 1{,}73 \\
1{,}06 & 1{,}18 & 1{,}26 & 1{,}46 & 1{,}92 & 2{,}58 & 3{,}74 & 6{,}36 & 4{,}20 & 2{,}96 & 2{,}18 & 1{,}66 & 1{,}26 & 1{,}10 & 1{,}07 & 0{,}98 & 1{,}38 & 1{,}42 \\
0{,}98 & 1{,}07 & 1{,}10 & 1{,}26 & 1{,}66 & 2{,}18 & 2{,}96 & 4{,}20 & 6{,}36 & 3{,}74 & 2{,}58 & 1{,}92 & 1{,}46 & 1{,}26 & 1{,}18 & 1{,}06 & 1{,}42 & 1{,}38 \\
1{,}07 & 1{,}10 & 1{,}07 & 1{,}19 & 1{,}54 & 1{,}97 & 2{,}47 & 2{,}91 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}58 & 1{,}92 & 1{,}65 & 1{,}53 & 1{,}34 & 1{,}73 & 1{,}59 \\
1{,}10 & 1{,}07 & 0{,}98 & 1{,}06 & 1{,}34 & 1{,}65 & 1{,}96 & 2{,}17 & 2{,}58 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}58 & 2{,}17 & 1{,}96 & 1{,}65 & 2{,}04 & 1{,}74 \\
1{,}07 & 0{,}98 & 0{,}86 & 0{,}89 & 1{,}12 & 1{,}34 & 1{,}53 & 1{,}65 & 1{,}92 & 2{,}58 & 3{,}74 & 6{,}36 & 3{,}74 & 2{,}91 & 2{,}47 & 1{,}97 & 2{,}27 & 1{,}80 \\
0{,}98 & 0{,}86 & 0{,}71 & 0{,}72 & 0{,}89 & 1{,}06 & 1{,}18 & 1{,}26 & 1{,}46 & 1{,}92 & 2{,}58 & 3{,}74 & 6{,}36 & 4{,}20 & 2{,}96 & 2{,}18 & 2{,}36 & 1{,}75 \\
1{,}06 & 0{,}89 & 0{,}72 & 0{,}71 & 0{,}86 & 0{,}98 & 1{,}07 & 1{,}10 & 1{,}26 & 1{,}66 & 2{,}18 & 2{,}96 & 4{,}20 & 6{,}36 & 3{,}74 & 2{,}58 & 2{,}69 & 1{,}92 \\
1{,}34 & 1{,}12 & 0{,}89 & 0{,}86 & 0{,}98 & 1{,}07 & 1{,}10 & 1{,}07 & 1{,}19 & 1{,}54 & 1{,}97 & 2{,}47 & 2{,}91 & 3{,}74 & 6{,}36 & 3{,}74 & 3{,}57 & 2{,}47 \\
1{,}65 & 1{,}34 & 1{,}06 & 0{,}98 & 1{,}07 & 1{,}10 & 1{,}07 & 0{,}98 & 1{,}06 & 1{,}34 & 1{,}65 & 1{,}96 & 2{,}17 & 2{,}58 & 3{,}74 & 6{,}36 & 5{,}05 & 3{,}15 \\
2{,}21 & 1{,}74 & 1{,}36 & 1{,}23 & 1{,}27 & 1{,}23 & 1{,}13 & 0{,}98 & 1{,}01 & 1{,}23 & 1{,}44 & 1{,}60 & 1{,}67 & 1{,}90 & 2{,}53 & 3{,}52 & 8{,}21 & 4{,}50 \\
3{,}52 & 2{,}53 & 1{,}90 & 1{,}67 & 1{,}60 & 1{,}44 & 1{,}23 & 1{,}01 & 0{,}98 & 1{,}13 & 1{,}23 & 1{,}27 & 1{,}23 & 1{,}36 & 1{,}74 & 2{,}21 & 4{,}50 & 8{,}21
\end{bmatrix} \tag{10.40}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\mathbb{H} = 10^{-2} \cdot \overbrace{\text{Nieznane } T}^{} \qquad \overbrace{\text{Nieznane } T}^{}$$

$$\begin{bmatrix}
-50 & 0{,}00 & 0{,}00 & 6{,}06 & 4{,}68 & 3{,}20 & 2{,}17 & 1{,}51 & 2{,}74 & 3{,}05 & 3{,}17 & 3{,}05 & 2{,}74 & 1{,}51 & 2{,}17 & 3{,}20 & 1{,}38 & 9{,}36 \\
0{,}00 & -50 & 0{,}00 & 9{,}36 & 5{,}40 & 2{,}86 & 1{,}67 & 1{,}07 & 3{,}05 & 3{,}17 & 3{,}05 & 2{,}74 & 2{,}34 & 1{,}72 & 2{,}28 & 3{,}02 & 2{,}21 & 6{,}06 \\
0{,}00 & 0{,}00 & -50 & 17{,}6 & 3{,}48 & 1{,}27 & 0{,}65 & 0{,}39 & 3{,}17 & 3{,}05 & 2{,}74 & 2{,}34 & 1{,}94 & 1{,}77 & 2{,}21 & 2{,}70 & 2{,}23 & 4{,}43 \\
1{,}27 & 3{,}48 & 17{,}6 & -50 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 3{,}48 & 3{,}18 & 2{,}70 & 2{,}21 & 1{,}77 & 1{,}94 & 2{,}34 & 2{,}74 & 2{,}66 & 4{,}61 \\
2{,}86 & 5{,}40 & 9{,}36 & 0{,}00 & -50 & 0{,}00 & 0{,}00 & 0{,}00 & 4{,}43 & 3{,}83 & 3{,}02 & 2{,}28 & 1{,}72 & 2{,}34 & 2{,}74 & 3{,}05 & 3{,}57 & 5{,}40 \\
3{,}20 & 4{,}68 & 6{,}06 & 0{,}00 & 0{,}00 & -50 & 0{,}00 & 0{,}00 & 6{,}06 & 4{,}68 & 3{,}20 & 2{,}17 & 1{,}51 & 2{,}74 & 3{,}05 & 3{,}17 & 4{,}12 & 5{,}35 \\
3{,}02 & 3{,}83 & 4{,}43 & 0{,}00 & 0{,}00 & 0{,}00 & -50 & 0{,}00 & 9{,}36 & 5{,}40 & 2{,}86 & 1{,}67 & 1{,}07 & 3{,}05 & 3{,}17 & 3{,}05 & 4{,}25 & 4{,}83 \\
2{,}70 & 3{,}18 & 3{,}48 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & -50 & 17{,}6 & 3{,}48 & 1{,}27 & 0{,}65 & 0{,}39 & 3{,}17 & 3{,}05 & 2{,}74 & 4{,}06 & 4{,}20 \\
2{,}74 & 3{,}05 & 3{,}17 & 0{,}39 & 0{,}65 & 1{,}27 & 3{,}48 & 17{,}6 & -50 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 3{,}48 & 3{,}18 & 2{,}70 & 4{,}20 & 4{,}06 \\
3{,}05 & 3{,}17 & 3{,}05 & 1{,}07 & 1{,}67 & 2{,}86 & 5{,}40 & 9{,}36 & 0{,}00 & -50 & 0{,}00 & 0{,}00 & 0{,}00 & 4{,}43 & 3{,}83 & 3{,}02 & 4{,}83 & 4{,}25 \\
3{,}17 & 3{,}05 & 2{,}74 & 1{,}51 & 2{,}17 & 3{,}20 & 4{,}68 & 6{,}06 & 0{,}00 & 0{,}00 & -50 & 0{,}00 & 0{,}00 & 6{,}06 & 4{,}68 & 3{,}20 & 5{,}35 & 4{,}12 \\
3{,}05 & 2{,}74 & 2{,}34 & 1{,}72 & 2{,}28 & 3{,}02 & 3{,}83 & 4{,}43 & 0{,}00 & 0{,}00 & 0{,}00 & -50 & 0{,}00 & 9{,}36 & 5{,}40 & 2{,}86 & 5{,}40 & 3{,}57 \\
2{,}74 & 2{,}34 & 1{,}94 & 1{,}77 & 2{,}21 & 2{,}70 & 3{,}18 & 3{,}48 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & -50 & 17{,}6 & 3{,}48 & 1{,}27 & 4{,}61 & 2{,}66 \\
2{,}70 & 2{,}21 & 1{,}77 & 1{,}94 & 2{,}34 & 2{,}74 & 3{,}05 & 3{,}17 & 0{,}39 & 0{,}65 & 1{,}27 & 3{,}48 & 17{,}6 & -50 & 0{,}00 & 0{,}00 & 4{,}43 & 2{,}23 \\
3{,}02 & 2{,}28 & 1{,}72 & 2{,}34 & 2{,}74 & 3{,}05 & 3{,}17 & 3{,}05 & 1{,}07 & 1{,}67 & 2{,}86 & 5{,}40 & 9{,}36 & 0{,}00 & -50 & 0{,}00 & 6{,}06 & 2{,}21 \\
3{,}20 & 2{,}17 & 1{,}51 & 2{,}74 & 3{,}05 & 3{,}17 & 3{,}05 & 2{,}74 & 1{,}51 & 2{,}17 & 3{,}20 & 4{,}68 & 6{,}06 & 0{,}00 & 0{,}00 & -50 & 9{,}36 & 1{,}38 \\
3{,}90 & 2{,}16 & 1{,}32 & 3{,}36 & 3{,}52 & 3{,}36 & 2{,}95 & 2{,}45 & 1{,}98 & 2{,}63 & 3{,}43 & 4{,}19 & 4{,}52 & 0{,}88 & 1{,}98 & 7{,}38 & -50 & 0{,}00 \\
7{,}38 & 1{,}98 & 0{,}88 & 4{,}52 & 4{,}19 & 3{,}43 & 2{,}63 & 1{,}98 & 2{,}45 & 2{,}95 & 3{,}36 & 3{,}52 & 3{,}36 & 1{,}32 & 2{,}16 & 3{,}90 & 0{,}00 & -50
\end{bmatrix} \tag{10.41}$$

Powtarzając powyżej wymienione kroki dla każdego elementu brzegowego, można wyznaczyć wartości w jednym wierszu macierzy. Zmieniając kolejno węzeł ξ_i, można uzyskać pełne macierze wpływu. Wyznaczoną macierz $\mathbb{G}$ przedstawia zależność (10.40). Ze względów technicznych związanych z rozmiarem macierzy ograniczono się to dwóch miejsc po przecinku, zaś wszystkie elementy macierzy zostały przemnożone przez 100. Podobnie wyznaczoną macierz $\mathbb{H}$ przedstawiono w zależności (10.41) i wartości jej elementów także odpowiednio przeskalowano, by polepszyć czytelność.

W zależności (10.40) zaznaczono kolumny od 9 do 13 oraz 17 i 18, które odpowiadają nieznanym wartościom strumieni ciepła q. Zaznaczone kolumny będą tworzyć część macierzy współczynników układu równań, który posłuży do wyznaczenie wszystkich nieznanych wartości temperatury i strumieni ciepła na brzegu.

Widać też wyraźnie istotną cechę macierzy wpływu metody elementów brzegowych – nie są one symetryczne. Wynika to bezpośrednio z tego, że wartości g_{ij} zależą nie tylko od odległości między punktem obserwacji ξ_i a j-tym elementem, ale także od kształtu i położenia tego ostatniego. Podobnie jak to ma miejsce w przypadku macie-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

rzy $\mathbb{G}$, opisującej strumienie ciepła, w zależności (10.41) zaznaczono kolumny dotyczące nieznanych wartości temperatury na brzegu. Jak widać z zależności (10.39), są to kolumny od 1 do 8 oraz od 14 do 16. Kolumny te zostaną wykorzystane do budowy macierzy współczynników $\mathbb{A}$ układu równań potrzebnego do wyznaczenia nieznanych wartości na brzegu obszaru. Należy zwrócić uwagę na to, że w macierzy $\mathbb{H}$ uwzględniono wartości równe $1/2$ leżące na głównej przekątnej, które w równaniu (10.33) zostały ujęte w wyrazie $\delta_{ij}B_i$.

Budując macierz $\mathbb{A}$, należy pamiętać o tym, że macierz $\mathbb{G}$ jest po lewej stronie układu równań (10.33), zaś macierz $\mathbb{H}$ po jego prawej stronie. W związku z tym przenoszone do macierzy $\mathbb{A}$ elementy macierzy $\mathbb{G}$ nie zmieniają znaku, zaś elementy macierzy $\mathbb{H}$ znak zmieniają. Zostało to zaznaczone nad odpowiednimi kolumnami macierzy $\mathbb{A}$ w zależności (10.42).

$$\mathbb{A} = 10^{-2}\cdot \begin{bmatrix} \overbrace{\phantom{50\ -0{,}0\ -0{,}0\ -6{,}1\ -4{,}7\ -3{,}2\ -2{,}2\ -1{,}5}}^{-h_{ij}} & \overbrace{\phantom{1{,}0\ 1{,}1\ 1{,}1\ 1{,}1\ 1{,}0}}^{g_{ij}} & \overbrace{\phantom{-1{,}5\ -2{,}2\ -3{,}2}}^{-h_{ij}} & \overbrace{\phantom{3{,}1\ 5{,}1}}^{g_{ij}} \\ 50 & -0{,}0 & -0{,}0 & -6{,}1 & -4{,}7 & -3{,}2 & -2{,}2 & -1{,}5 & 1{,}0 & 1{,}1 & 1{,}1 & 1{,}1 & 1{,}0 & -1{,}5 & -2{,}2 & -3{,}2 & 3{,}1 & 5{,}1 \\ -0{,}0 & 50 & -0{,}0 & -9{,}4 & -5{,}4 & -2{,}9 & -1{,}7 & -1{,}1 & 1{,}1 & 1{,}1 & 1{,}1 & 1{,}0 & 0{,}9 & -1{,}7 & -2{,}3 & -3{,}0 & 2{,}5 & 3{,}6 \\ -0{,}0 & -0{,}0 & 50 & -17{,}6 & -3{,}5 & -1{,}3 & -0{,}6 & -0{,}4 & 1{,}1 & 1{,}1 & 1{,}0 & 0{,}9 & 0{,}7 & -1{,}8 & -2{,}2 & -2{,}7 & 1{,}9 & 2{,}7 \\ -1{,}3 & -3{,}5 & -17{,}6 & 50 & -0{,}0 & -0{,}0 & -0{,}0 & -0{,}0 & 1{,}3 & 1{,}2 & 1{,}1 & 0{,}9 & 0{,}7 & -1{,}9 & -2{,}3 & -2{,}7 & 1{,}7 & 2{,}4 \\ -2{,}9 & -5{,}4 & -9{,}4 & -0{,}0 & 50 & -0{,}0 & -0{,}0 & -0{,}0 & 1{,}6 & 1{,}5 & 1{,}3 & 1{,}1 & 0{,}9 & -2{,}3 & -2{,}7 & -3{,}1 & 1{,}8 & 2{,}3 \\ -3{,}2 & -4{,}7 & -6{,}1 & -0{,}0 & -0{,}0 & 50 & -0{,}0 & -0{,}0 & 2{,}2 & 2{,}0 & 1{,}7 & 1{,}3 & 1{,}1 & -2{,}7 & -3{,}1 & -3{,}2 & 1{,}7 & 2{,}0 \\ -3{,}0 & -3{,}8 & -4{,}4 & -0{,}0 & -0{,}0 & -0{,}0 & 50 & -0{,}0 & 2{,}9 & 2{,}5 & 2{,}0 & 1{,}5 & 1{,}2 & -3{,}1 & -3{,}2 & -3{,}1 & 1{,}6 & 1{,}7 \\ -2{,}7 & -3{,}2 & -3{,}5 & -0{,}0 & -0{,}0 & -0{,}0 & -0{,}0 & 50 & 4{,}2 & 3{,}0 & 2{,}2 & 1{,}7 & 1{,}3 & -3{,}2 & -3{,}1 & -2{,}7 & 1{,}4 & 1{,}4 \\ -2{,}7 & -3{,}1 & -3{,}2 & -0{,}4 & -0{,}6 & -1{,}3 & -3{,}5 & -17{,}6 & 6{,}4 & 3{,}7 & 2{,}6 & 1{,}9 & 1{,}5 & -3{,}5 & -3{,}2 & -2{,}7 & 1{,}4 & 1{,}4 \\ -3{,}1 & -3{,}2 & -3{,}1 & -1{,}1 & -1{,}7 & -2{,}9 & -5{,}4 & -9{,}4 & 3{,}7 & 6{,}4 & 3{,}7 & 2{,}6 & 1{,}9 & -4{,}4 & -3{,}8 & -3{,}0 & 1{,}7 & 1{,}6 \\ -3{,}2 & -3{,}1 & -2{,}7 & -1{,}5 & -2{,}2 & -3{,}2 & -4{,}7 & -6{,}1 & 2{,}6 & 3{,}7 & 6{,}4 & 3{,}7 & 2{,}6 & -6{,}1 & -4{,}7 & -3{,}2 & 2{,}0 & 1{,}7 \\ -3{,}1 & -2{,}7 & -2{,}3 & -1{,}7 & -2{,}3 & -3{,}0 & -3{,}8 & -4{,}4 & 1{,}9 & 2{,}6 & 3{,}7 & 6{,}4 & 3{,}7 & -9{,}4 & -5{,}4 & -2{,}9 & 2{,}3 & 1{,}8 \\ -2{,}7 & -2{,}3 & -1{,}9 & -1{,}8 & -2{,}2 & -2{,}7 & -3{,}2 & -3{,}5 & 1{,}5 & 1{,}9 & 2{,}6 & 3{,}7 & 6{,}4 & -17{,}6 & -3{,}5 & -1{,}3 & 2{,}4 & 1{,}7 \\ -2{,}7 & -2{,}2 & -1{,}8 & -1{,}9 & -2{,}3 & -2{,}7 & -3{,}1 & -3{,}2 & 1{,}3 & 1{,}7 & 2{,}2 & 3{,}0 & 4{,}2 & 50 & -0{,}0 & -0{,}0 & 2{,}7 & 1{,}9 \\ -3{,}0 & -2{,}3 & -1{,}7 & -2{,}3 & -2{,}7 & -3{,}1 & -3{,}2 & -3{,}1 & 1{,}2 & 1{,}5 & 2{,}0 & 2{,}5 & 2{,}9 & -0{,}0 & 50 & -0{,}0 & 3{,}6 & 2{,}5 \\ -3{,}2 & -2{,}2 & -1{,}5 & -2{,}7 & -3{,}1 & -3{,}2 & -3{,}1 & -2{,}7 & 1{,}1 & 1{,}3 & 1{,}7 & 2{,}0 & 2{,}2 & -0{,}0 & -0{,}0 & 50 & 5{,}1 & 3{,}1 \\ -3{,}9 & -2{,}2 & -1{,}3 & -3{,}4 & -3{,}5 & -3{,}4 & -2{,}9 & -2{,}5 & 1{,}0 & 1{,}2 & 1{,}4 & 1{,}6 & 1{,}7 & -0{,}9 & -2{,}0 & -7{,}4 & 8{,}2 & 4{,}5 \\ -7{,}4 & -2{,}0 & -0{,}9 & -4{,}5 & -4{,}2 & -3{,}4 & -2{,}6 & -2{,}0 & 1{,}0 & 1{,}1 & 1{,}2 & 1{,}3 & 1{,}2 & -1{,}3 & -2{,}2 & -3{,}9 & 4{,}5 & 8{,}2 \end{bmatrix} \tag{10.42}$$

Znane wartości temperatury i strumieni ciepła na brzegu określają kolumny macierzy $\mathbb{G}$ oraz $\mathbb{H}$ które posłużą do zbudowania macierzy $\mathbb{B}$. Macierz $\mathbb{B}$ pomnożona przez macierz $\mathbb{y}$ utworzy wektor kolumnowy, stanowiący prawą stronę układu równań

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

(10.34). Podobnie w przypadku przy macierzy $\mathbb{A}$, budując macierz $\mathbb{B}$, należy pamiętać, by odpowiednio zmieniać znaki elementów w przenoszonych kolumnach. Zatem elementom w kolumnach z macierzy $\mathbb{G}$ zmienia się znak, zaś elementy w kolumnach z macierzy $\mathbb{H}$ znaku nie zmieniają. Ta ważna informacja została zaznaczona w zależności (10.43) w nagłówkach odpowiednich kolumn.

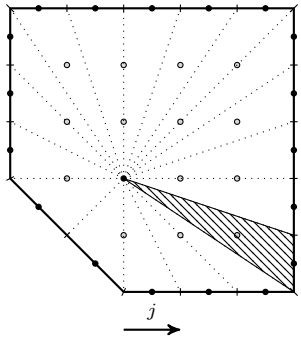
$$\mathbb{B} = 10^{-2}\cdot \quad \overbrace{\begin{matrix} -g_{ij} & & & & & & & & & & & & & & & & & \end{matrix}}^{} \quad \overbrace{\begin{matrix} h_{ij} & & & & & & & & & & & & & & & & & \end{matrix}}^{}$$

$$\begin{bmatrix}
-6{,}4 & -3{,}7 & -2{,}6 & -2{,}2 & -2{,}0 & -1{,}7 & -1{,}3 & -1{,}1 & 2{,}7 & 3{,}1 & 3{,}2 & 3{,}1 & 2{,}7 & -1{,}1 & -1{,}3 & -1{,}7 & 1{,}4 & 9{,}4 \\
-3{,}7 & -6{,}4 & -3{,}7 & -2{,}9 & -2{,}5 & -2{,}0 & -1{,}5 & -1{,}2 & 3{,}1 & 3{,}2 & 3{,}1 & 2{,}7 & 2{,}3 & -0{,}9 & -1{,}1 & -1{,}3 & 2{,}2 & 6{,}1 \\
-2{,}6 & -3{,}7 & -6{,}4 & -4{,}2 & -3{,}0 & -2{,}2 & -1{,}7 & -1{,}3 & 3{,}2 & 3{,}1 & 2{,}7 & 2{,}3 & 1{,}9 & -0{,}7 & -0{,}9 & -1{,}1 & 2{,}2 & 4{,}4 \\
-2{,}2 & -3{,}0 & -4{,}2 & -6{,}4 & -3{,}7 & -2{,}6 & -1{,}9 & -1{,}5 & 3{,}5 & 3{,}2 & 2{,}7 & 2{,}2 & 1{,}8 & -0{,}7 & -0{,}9 & -1{,}0 & 2{,}7 & 4{,}6 \\
-2{,}0 & -2{,}5 & -2{,}9 & -3{,}7 & -6{,}4 & -3{,}7 & -2{,}6 & -1{,}9 & 4{,}4 & 3{,}8 & 3{,}0 & 2{,}3 & 1{,}7 & -0{,}9 & -1{,}0 & -1{,}1 & 3{,}6 & 5{,}4 \\
-1{,}7 & -2{,}0 & -2{,}2 & -2{,}6 & -3{,}7 & -6{,}4 & -3{,}7 & -2{,}6 & 6{,}1 & 4{,}7 & 3{,}2 & 2{,}2 & 1{,}5 & -1{,}0 & -1{,}1 & -1{,}1 & 4{,}1 & 5{,}3 \\
-1{,}3 & -1{,}5 & -1{,}6 & -1{,}9 & -2{,}6 & -3{,}7 & -6{,}4 & -3{,}7 & 9{,}4 & 5{,}4 & 2{,}9 & 1{,}7 & 1{,}1 & -1{,}1 & -1{,}1 & -1{,}1 & 4{,}3 & 4{,}8 \\
-1{,}1 & -1{,}2 & -1{,}3 & -1{,}5 & -1{,}9 & -2{,}6 & -3{,}7 & -6{,}4 & 17{,}6 & 3{,}5 & 1{,}3 & 0{,}6 & 0{,}4 & -1{,}1 & -1{,}1 & -1{,}0 & 4{,}1 & 4{,}2 \\
-1{,}0 & -1{,}1 & -1{,}1 & -1{,}3 & -1{,}7 & -2{,}2 & -3{,}0 & -4{,}2 & -50 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}3 & -1{,}2 & -1{,}1 & 4{,}2 & 4{,}1 \\
-1{,}1 & -1{,}1 & -1{,}1 & -1{,}2 & -1{,}5 & -2{,}0 & -2{,}5 & -2{,}9 & 0{,}0 & -50 & 0{,}0 & 0{,}0 & 0{,}0 & -1{,}6 & -1{,}5 & -1{,}3 & 4{,}8 & 4{,}3 \\
-1{,}1 & -1{,}1 & -1{,}0 & -1{,}1 & -1{,}3 & -1{,}7 & -2{,}0 & -2{,}2 & 0{,}0 & 0{,}0 & -50 & 0{,}0 & 0{,}0 & -2{,}2 & -2{,}0 & -1{,}7 & 5{,}3 & 4{,}1 \\
-1{,}1 & -1{,}0 & -0{,}9 & -0{,}9 & -1{,}1 & -1{,}3 & -1{,}5 & -1{,}6 & 0{,}0 & 0{,}0 & 0{,}0 & -50 & 0{,}0 & -2{,}9 & -2{,}5 & -2{,}0 & 5{,}4 & 3{,}6 \\
-1{,}0 & -0{,}9 & -0{,}7 & -0{,}7 & -0{,}9 & -1{,}1 & -1{,}2 & -1{,}3 & 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 & -50 & -4{,}2 & -3{,}0 & -2{,}2 & 4{,}6 & 2{,}7 \\
-1{,}1 & -0{,}9 & -0{,}7 & -0{,}7 & -0{,}9 & -1{,}0 & -1{,}1 & -1{,}1 & 0{,}4 & 0{,}6 & 1{,}3 & 3{,}5 & 17{,}6 & -6{,}4 & -3{,}7 & -2{,}6 & 4{,}4 & 2{,}2 \\
-1{,}3 & -1{,}1 & -0{,}9 & -0{,}9 & -1{,}0 & -1{,}1 & -1{,}1 & -1{,}1 & 1{,}1 & 1{,}7 & 2{,}9 & 5{,}4 & 9{,}4 & -3{,}7 & -6{,}4 & -3{,}7 & 6{,}1 & 2{,}2 \\
-1{,}7 & -1{,}3 & -1{,}1 & -1{,}0 & -1{,}1 & -1{,}1 & -1{,}1 & -1{,}0 & 1{,}5 & 2{,}2 & 3{,}2 & 4{,}7 & 6{,}1 & -2{,}6 & -3{,}7 & -6{,}4 & 9{,}4 & 1{,}4 \\
-2{,}2 & -1{,}7 & -1{,}4 & -1{,}2 & -1{,}3 & -1{,}2 & -1{,}1 & -1{,}0 & 2{,}0 & 2{,}6 & 3{,}4 & 4{,}2 & 4{,}5 & -1{,}9 & -2{,}5 & -3{,}5 & -50 & 0{,}0 \\
-3{,}5 & -2{,}5 & -1{,}9 & -1{,}7 & -1{,}6 & -1{,}4 & -1{,}2 & -1{,}0 & 2{,}5 & 2{,}9 & 3{,}4 & 3{,}5 & 3{,}4 & -1{,}4 & -1{,}7 & -2{,}2 & 0{,}0 & -50{,}0
\end{bmatrix} \tag{10.43}$$

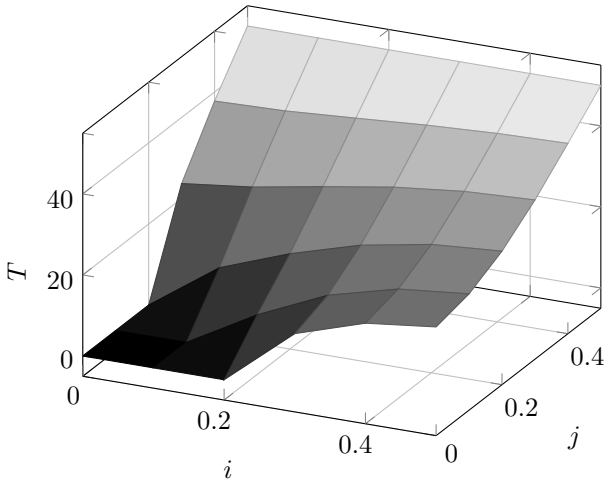
Po wyznaczeniu wszystkich wartości temperatury i strumieni ciepła na brzegu można przystąpić do wyznaczenia wartości temperatury wewnątrz analizowanego obszaru na podstawie równania (10.37). Aby wyznaczyć wartość temperatury w punkcie wewnątrz analizowanego obszaru, konieczne jest dokonanie odpowiednich całkowań. Postępowanie przy wyznaczaniu temperatury w punktach wewnętrznych ilustruje rysunek 10.6. Dla każdego z punktów należy zrealizować całkowanie po całym obwodzie obszaru Ω. Wyniki obliczeń przedstawiono w postaci wykresu przestrzennego na rysunku 10.7.

Węzły na brzegu obszaru i wewnątrz nie tworzyły regularnej siatki.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 10.6. Całkowanie na potrzeby wyznaczania temperatury wewnątrz obszaru



Rysunek 10.7. Rozkład temperatury wyznaczony za pomocą MEB

Z tego powodu siatka użyta w wizualizacji zawiera węzły wewnętrzne oraz punkty leżące na granicach elementów brzegowych. Wartości na krańcach elementów wyznaczono jako średnią z temperatury sąsiednich elementów.

10.5. Podsumowanie

Niniejszy rozdział został pomyślany jako pewna prezentacja idei metody elementów brzegowych. Z tego powodu nie stanowi wyczerpującego wykładu tak niezwykle bogatej w odmiany metody. Wykorzystanie równania opisującego ustalony przepływ ciepła pozwoliło na zaprezentowanie tematyki w nieco mniej abstrakcyjny sposób. Osoby pragnące pogłębić wiedzę na temat metody elementów brzegowych powinny zapoznać się z pozycjami bibliograficznymi [21, 25, 32, 33].

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 11

Metoda elementów skończonych

11.1. Wstęp

Metoda elementów skończonych (*Finite Element Method*) MES jest obecnie jedną z najczęściej stosowanych metod obliczeniowych w nauce i technice. Przez ostatnie prawie prawie lat doczekała się wielu publikacji dokładnie ją omawiających. Wprowadzać pojęcia metody elementów skończonych można na różne sposoby, jednym z nich jest sformułowanie wariacyjne, innym metoda Galerkina. Taki sposób przedstawiania tej metody wymaga dobrego przygotowania matematycznego – często znacznie lepszego niż posiadane przez inżyniera. Z tego też powodu zagadnienia przedstawione w niniejszym rozdziale będą znacznie uproszczone, pozwalające jednak zrozumieć istotę metody oraz jej wady i zalety. Jako jedna z podstawowych metod obliczeniowych do rozwiązywania problemów inżynierskich, warta jest krótkiego przybliżenia.

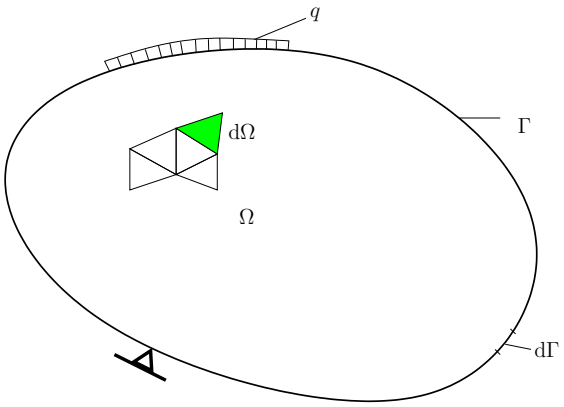
11.2. Idea metody elementów skończonych

Metoda elementów skończonych służy do rozwiązywania problemów brzegowych i początkowo-brzegowych. Określa się obszar rozwiązania Ω, brzeg Γ z warunkami brzegowymi, element objętości obszaru $d\Omega$ i element brzegu $d\Gamma$, tak jak pokazano na rysunku 11.1. Na brzegu mogą być zdefiniowane obciążenia, można uwzględniać siły masowe, które działają na cały obszar rozwiązania.

Aby rozwiązać problem, dzieli się obszar, w którym poszukuje się rozwiązania, na podobszary o znanym i nieskomplikowanym kształcie. Owe podobszary, nazywane elementami skończonymi, są określane przez węzły stanowiące ich wierzchołki. Wierzchołki łączone są krawędziami. Rozkład szukanej wielkości wewnątrz elementu jest interpolowany za pomocą tzw. funkcji kształtu. Funkcje kształtu muszą spełnić przynajmniej niektóre z wymienionych poniżej podstawowych warunków zapewniających zbieżność rozwiązania. Zatem funkcje kształtu muszą:

1. zapewniać ciągłość przemieszczeń wewnątrz elementu, oraz ich zgodność na granicach elementów,
2. umożliwić opisanie stałych przemieszczeń elementu, co oznacza ruch elementu jako ciała sztywnego,
3. umożliwić opisanie stałych odkształceń wewnątrz elementu, występujących przy odpowiednich przemieszczeniach węzłów.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



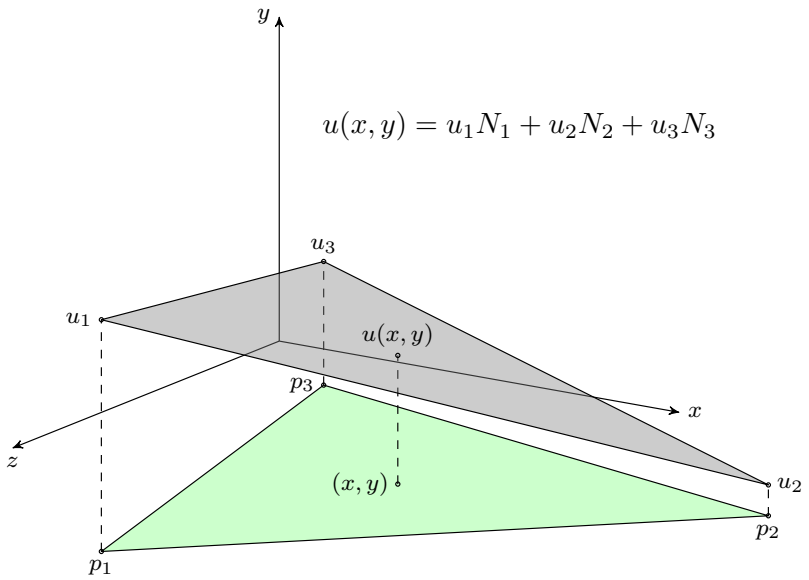
Rysunek 11.1. Schemat obszaru rozwiązania dla MES

Elementy spełniające wszystkie wymienione wcześniej warunki są zwane elementami zgodnymi, czyli dostosowanymi, zaś elementy spełniające tylko drugi i trzeci warunek nazywa się elementami zupełnymi, czyli niedostosowanymi. Rozwiązanie problemu zdyskretyzowanego za pomocą elementów zgodnych jest zbieżne do rozwiązania dokładnego od dołu. Gdy zostaną użyte elementy zupełne, rozwiązanie nie zawsze jest zbieżne, a gdy już jest zbieżne, to jest wolniej zbieżne niż przy elementach zgodnych. Co więcej, takie rozwiązanie może być zbieżne do prawidłowego rozwiązania od góry lub od dołu.

Jak wspomniano, rozkład analizowanej wielkości w elemencie jest określany przez interpolację funkcjami kształtu na podstawie wartości w węzłach, jak to pokazano na rysunku 11.2. Zatem warto pamiętać, że metoda pozwala na wyznaczanie wartości rozwiązania równania różniczkowego w węzłach. Z tego wynika, że w miejscach, w których następuje szybka zmiana rozwiązania, konieczne jest stosowanie gęstszej siatki, by właściwie odwzorować przebieg rozwiązania.

Ogólny tok postępowania rozwiązywania problemu opisanego równaniem różniczkowym cząstkowym (r. r. cz) za pomocą metody elementów skończonych przedstawiono w postaci pseudokodu w algorytmie 11.1. W następnych podrozdziałach zostaną omówione poszczególne etapy tego algorytmu. Dla problemów wyznaczania odkształceń i naprężeń można sformułować równania metody elementów skończonych na różne sposoby [43]. Jeden z nich polega na bezpośrednim wyznaczaniu przemieszczeń węzłów przez rozwiązanie odpowiedniego układu równań. Nieco inne podejście polega na wyznaczaniu naprężeń i sił węzłowych. Oczywiście i w jednym i w drugim przypadku można wyznaczyć zarówno przemieszczenia, jak i naprężenia. Znane są sformułowania pozwalające wyznaczać równocześnie naprężenia i odkształcenia, na przykład funkcjonał Hu-Washitzu daje model przemieszczeniowo-odkształceniowo-naprężeniowy. Należy pamiętać, iż wielkości bezpośrednio wyznaczane na podstawie rozwiązania odpowiedniego układu równań są dokładniejsze niż wielkości wyznaczane na ich podstawie na przykład jako ich pochodne.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.2. Określanie wartości wewnątrz elementu

Algorytm 11.1. Metoda elementów skończonych

Require: problem w postaci r. r. cz
Require: warunki brzegowe
Ensure: postawiony właściwie
  podział obszaru rozwiązania na elementy skończone
  for all elementy obszaru **do**
    wyznaczyć macierze opisujące właściwości elementów
    agregować macierze elementów do macierzy globalnych
  end for
  for all elementy na brzegu **do**
    odebrać odpowiednie stopnie swobody
    nałożyć warunki Neumana
    nałożyć warunki mieszane
  end for
  rozwiązać uzyskany układ równań algebraicznych
  wizualizacja rozwiązania

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

11.3. Matematyczny opis układów mechanicznych

Problemy mechaniki ośrodków ciągłych opisywane są równaniami różniczkowymi cząstkowymi z podanymi warunkami brzegowymi oraz początkowymi. W praktyce wygodniej jest przekształcić problem opisany równaniem różniczkowym do równoważnego problemu wariacyjnego. Rozwiązanie problemu sformułowanego wariacyjnie sprowadza się do wyznaczenia funkcji, dla której pewien funkcjonał osiąga minimum. Znalezienie dokładnej postaci funkcji stanowiącej rozwiązanie zazwyczaj nastręcza znacznych trudności, stąd też opracowano metody przybliżone rozwiązywania takich problemów. Tradycyjnie mówi się o metodzie Rayleigha–Ritza oraz o metodach residuów ważonych. Spośród metod residuów ważonych najczęściej stosowana jest metoda Galerkina. W metodzie Rayleigha–Ritza niewiadome parametry funkcji aproksymujących są wyznaczane z warunku stacjonarności funkcjonału, zaś w metodzie Galerkina wyznacza się te współczynniki z warunku ortogonalności. W następnych podrozdziałach krótko omówione zostaną obie te metody – metoda Ritza oraz Galerkina.

11.4. Przybliżone metody rozwiązywania równań

Jak już wcześniej stwierdzono, układy fizyczne, a w szczególności układy mechaniczne, są opisywane równaniami różniczkowymi, często równaniami różniczkowymi cząstkowymi. Rozwiązanie równania różniczkowego cząstkowego polega na znalezieniu funkcji, która to równanie spełni na brzegu i całym obszarze rozwiązania dla wszystkich chwil czasowych. Najczęściej można to zrobić w prostych geometrycznie przypadkach i nieskomplikowanych równaniach. Od lat takie rozwiązania analityczne służyły do rozwiązywania problemów technicznych, pozwalały na projektowanie części maszyn i urządzeń spełniających swoje zadania. Nie pozwalały natomiast na przeprowadzanie działań optymalizacyjnych, służących zmniejszaniu materiałochłonności czy masy urządzeń. Dlatego do analizy układów mechanicznych zaczęto stosować metody przybliżone rozwiązywania równań różniczkowych cząstkowych. Dwie wybrane metody zostaną poniżej pokrótce omówione.

11.4.1. Metoda Rayleigha–Ritza

Metoda Rayleigha–Ritza polega na szukaniu rozwiązania równania różniczkowego przez minimalizację pewnego funkcjonału. Dlatego to podejście jest nazywane sformułowaniem wariacyjnym. Aby można było zastosować tę metodę, konieczne jest znalezienie takiego funkcjonału, którego minimum daje rozwiązanie równania różniczkowego. Nie dla każdego równania znana jest postać funkcjonału. Zatem nie jest to metoda ogólna. Zazwyczaj dla układów mechanicznych można znaleźć odpowiednie funkcjonały, ponieważ stan równowagi układu mechanicznego opisuje konfiguracja o najmniejszej możliwej energii potencjalnej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rachunek wariacyjny zajmuje się minimalizacją lub maksymalizacją funkcjonałów. Funkcjonały są to takie funkcje, których argumentami są funkcje. Często funkcjonały są wyrażane w postaci całek oznaczonych z wyrażeń zawierających funkcje i ich pochodne, a ich wartość zależy od drogi całkowania wyrażenia podcałkowego. Przykładem sformułowania wariacyjnego może być zasada Fermata, która mówi, że promień świetlny wybierze najkrótszą możliwą drogę optyczną, zależną od parametrów ośrodka. Podobnie w mechanice zdefiniowano zasadę najmniejszego działania wyrażoną na przykład zasadą Hamiltona.

Metoda Rayleigha–Ritza polega na minimalizowaniu całki z funkcjonału F odpowiadającego rozwiązywanemu problemowi, co można zapisać

$$L = \int_{x_1}^{x_2} F\left(x, w(x), \frac{\mathrm{d}w(x)}{\mathrm{d}x}\right) \mathrm{d}x$$

Punkt stacjonarny funkcjonału jest rozwiązaniem problemu sformułowanego wariacyjnie.

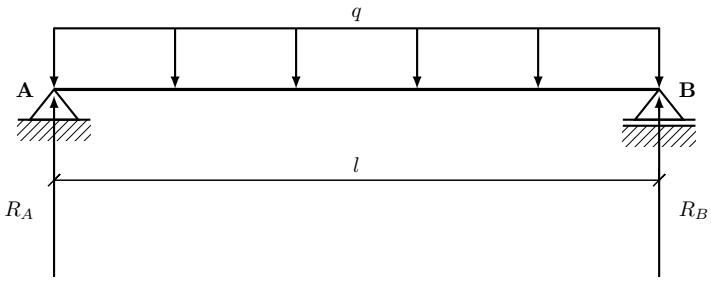
Przykład obliczeniowy

Sformułowanie problemu. Rozwiązać równanie różniczkowe opisujące linię ugięcia belki przedstawionej na rysunku 11.3, wykorzystując metodę Rayleigha–Ritza.

Dane. Znana jest długość belki l, moduł Younga E, moment bezwładności przekrojów belki względem osi prostopadłej do rysunku I oraz wartość obciążenia q. Wartości poszczególnych wielkości zestawiono poniżej.

Wielkość	Wartość	Jednostka
b	0,1	[m]
h	0,1	[m]
l	1	[m]
q	2000	[N/m]
E	$2{,}1 \cdot 10^9$	[Pa]
$I = bh^3/12$	$8{,}333 \cdot 10^{-6}$	[m^4]

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.3. Schemat obciążenia rozwiązywanej belki

Rozwiązanie

Wyznaczenie reakcji podpór. Aby określić wartości sił działających na belkę, należy uwolnić ją od więzów, zaś w ich miejsce przyłożyć siły reakcji o takich wartościach, by belka pozostała w stanie równowagi. Reakcje podpór można wyznaczyć z równań statyki, na przykład suma rzutów wszystkich sił na oś Y ma być zerem i suma momentów sił względem punktu podparcia A ma być zerem, co daje układ równań (11.1).

$$\begin{cases} \sum F_y = R_A - ql + R_B = 0 \\ \sum M_A = -ql\frac{l}{2} + R_B l = 0 \end{cases} \tag{11.1}$$

Rozwiązując układ równań (11.1), można wyznaczyć wartości reakcji jako

$$R_A = R_B = \frac{ql}{2}$$

Wobec tego momenty zginające w przekrojach poprzecznych belki mogą być opisane następującą zależnością

$$M(x) = R_A x - \frac{qx^2}{2} = 1000\left(x - x^2\right) \tag{11.2}$$

Linia ugięcia belki $w(x)$ przy założeniu niewielkich odkształceń może być opisana równaniem

$$EI\frac{\mathrm{d}^2 w(x)}{\mathrm{d}x^2} + M(x) = 0 \tag{11.3}$$

Warunki brzegowe wynikają ze sposobu podparcia belki, i są opisane równaniami (11.4).

$$\begin{aligned} x &= 0 \qquad & w &= 0 \\ x &= l \qquad & w &= 0 \end{aligned} \tag{11.4}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po przekształceniu równanie ugięcia belki można zapisać jako

$$\frac{\mathrm{d}^2 w}{\mathrm{d}x^2} = -\frac{R_A - q\frac{x^2}{2}}{EI}$$

Łatwo można sprawdzić, że równanie linii ugięcia wygląda następująco

$$\frac{\mathrm{d}^2 w}{\mathrm{d}x^2} + \frac{2}{35}\left(x - x^2\right) = 0 \tag{11.5}$$

Warunki brzegowe $w(0) = 0$ i $w(l) = 0$ wynikają ze sposobu podparcia belki. Aby zweryfikować dokładność zastosowanej metody przybliżonej, warto wyznaczyć rozwiązanie dokładne tego równania. Po dwukrotnym scałkowaniu i wykorzystaniu warunków brzegowych, można stwierdzić, że linia ugięcia analizowanej belki może być opisana wielomianem stopnia czwartego przedstawionym w zależności (11.6).

$$w(x) = \frac{x^4 - 2x^3 + x}{210} \tag{11.6}$$

Bez wkładania większego wysiłku można sprawdzić, że funkcjonał odpowiadający rozwiązywanemu równaniu ma następującą postać

$$F = -\frac{1}{2}\left(\frac{\mathrm{d}w(x)}{\mathrm{d}x}\right)^2 + \frac{2}{35}\left(x - x^2\right) w(x)$$

Poniżej zostanie zademonstrowany sposób przybliżonego rozwiązania równania (11.5) z wykorzystaniem jednej oraz dwóch funkcji próbnych spełniających warunki brzegowe.

Jedna funkcja próbna wykorzystana do rozwiązania równania wybrana została arbitralnie i ma postać

$$N_1 = x\left(1 - x^2\right), \quad \text{czyli} \quad \tilde{w}(x, a_1) = a_1 x\left(1 - x^2\right)$$

Zatem całka z funkcjonału przedstawia się zależnością (11.7) i wynosi

$$\begin{aligned} L &= \int_0^1 -\frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(a_1\, x\, \left(1 - x^2\right)\right)\right)^2 + \frac{2}{35}\, a_1\, x\, \left(1 - x^2\right)\, \left(x - x^2\right) \mathrm{d}x = \\ &= \int_0^1 -\frac{1}{2}\left(a_1\, \left(1 - x^2\right) - 2\, a_1\, x^2\right)^2 + \frac{2}{35} a_1\, x\, \left(1 - x^2\right)\, \left(x - x^2\right) \mathrm{d}x = \\ &= -\frac{140\, a_1^2 - a_1}{350} \end{aligned} \tag{11.7}$$

Minimum takiej całki (11.7) można wyznaczyć, obliczając odpowiednie pochodne

$$\frac{\mathrm{d}L}{\mathrm{d}x} = 0 \tag{11.8}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\frac{\mathrm{d}L}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{140\,a_1^2 - a_1}{350}\right) = \frac{1 - 280a_1}{350} = 0 \tag{11.9}$$

Zatem poszukiwany współczynnik a_1 przyjmuje wartość

$$a_1 = \frac{1}{280}$$

Wobec tego

$$\tilde{w}(x) = \frac{1}{280}x(1 - x^2)$$

stanowi przybliżone rozwiązanie równania różniczkowego opisującego linię ugięcia belki za pomocą metody Rayleigha–Ritza, gdy przyjmie się jedną funkcję próbną $N_1(x)$.

Dwie funkcje próbne wykorzystane do rozwiązania równania zostały przyjęte następująco

$$N_1 = x(1 - x^2), \quad N_2 = x(1 - x^4)$$

Przyjęte przybliżone rozwiązanie ma postać

$$\tilde{w}(x) = a_1N_1 + a_2N_2 = a_1x(1 - x^2) + a_2x(1 - x^4)$$

Podobnie jak przy rozwiązywaniu równania za pomocą przybliżonego rozwiązania składającego się z jednej funkcji próbnej, należy obliczyć całkę z funkcjonału, a następnie wyznaczyć warunki jego stacjonarności. Osiąga się to, przyrównując poszczególne pochodne cząstkowe do zera. W ten sposób uzyskuje się układ równań pozwalający wyznaczyć poszukiwane wartości współczynników a_1 i a_2. Całka z funkcjonału przyjmuje postać (11.10)

$$\begin{aligned} L = \int_0^1 -\frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(a_1\,x\,\left(1 - x^2\right) + a_2\,x\,\left(1 - x^4\right)\right)\right)^2 + \\ + \frac{2}{35}\left(x - x^2\right)\,\left(a_1\,x\,\left(1 - x^2\right) + a_2\,x\,\left(1 - x^4\right)\right)\mathrm{d}x = \\ = \frac{126\,a_1 - 17\,640\,a_1^2 - (50\,400\,a_1 - 165)\,a_2 - 39\,200\,a_2^2}{44\,100} \end{aligned} \tag{11.10}$$

Jak wspomniano, układ równań do wyznaczenia współczynników uzyskuje się z warunku

$$\frac{\partial L}{\partial a_i} = 0.$$

Zatem odpowiednie równania mają postać

$$\left\{\begin{array}{rcl} \frac{\partial L}{\partial a_1} & = & 0 \\ \frac{\partial L}{\partial a_2} & = & 0 \end{array}\right. \Rightarrow \left\{\begin{array}{rcl} \frac{126 - 35\,280\,a_1 - 50\,400\,a_2}{44\,100} & = & 0 \\ \frac{165 - 50\,400\,a_1 - 78\,400\,a_2}{44\,100} & = & 0 \end{array}\right. \tag{11.11}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Uzyskane współczynniki funkcji próbnych, określające rozwiązanie najbardziej zbliżone do dokładnego mają wartość

$$a_1 = \frac{31}{4480}, \quad a_2 = -\frac{3}{1280}$$

Jak wcześniej wspomniano, podstawową trudnością w stosowaniu metody Rayleigha–Ritza jest wyznaczenie odpowiedniego funkcjonału. Nie zawsze taki funkcjonał łatwo jest określić. Z tego też powodu znacznie częściej używa się metody Galerkina, która nie wymaga takiego funkcjonału, a daje podobne wyniki.

11.4.2. Metoda Galerkina

Nazwa tej metody pochodzi od nazwiska Borisa Grigorjewicza Galerkina (1891--1945), jest przedstawicielką grupy metod zwanych metodami reszt (residuów) ważonych. Metoda Galerkina jest metodą przybliżoną rozwiązywania problemów brzegowych. Zostanie teraz przedstawiona na przykładzie często spotykanego w mechanice równania różniczkowego w postaci

$$\mathrm{D}y - f = 0 \tag{11.12}$$

gdzie D jest liniowym operatorem różniczkowym, y jest szukaną funkcją, zaś f jest dane. Niech rozwiązanie da się przybliżyć kombinacją liniową funkcji aproksymacyjnych ϕ_i, które podobnie jak w metodzie Rayleigha–Ritza są nazywane funkcjami próbnymi

$$y \approx \tilde{y} = \sum_{i=1}^{n} c_i \phi_i \tag{11.13}$$

Aby powyższa zależność mogła być spełniona, wszystkie funkcje aproksymujące muszą spełniać warunki brzegowe. Zatem równanie (11.12) przyjmie postać

$$\mathrm{D}\tilde{y} - f \neq 0$$

co można zapisać

$$\mathrm{D}\tilde{y} - f = R \tag{11.14}$$

gdzie R jest nazywane residuum i stanowi błąd rozwiązania. Należy pamiętać o tym, że R jest funkcją współczynników c_i oraz zmiennych przestrzennych. Im mniejszą wartość przyjmuje residuum w całym przedziale, tym $\tilde{y}$ stanowi dokładniejsze przybliżenie rozwiązania. Dla danych funkcji próbnych wartość reszty R zależy od wyboru współczynników c_i. Aby wyznaczyć współczynniki c_i, wprowadza się funkcje wagowe w_i, tak dobrane by

$$\int_{\Omega} (\mathrm{D}\tilde{y} - f) w_i \mathrm{d}\Omega = \int_{\Omega} R w_i \mathrm{d}\Omega = 0, \qquad \text{dla } i = 1, 2, \ldots, n \tag{11.15}$$

W metodzie Galerkina przyjmuje się funkcje wagowe identyczne z funkcjami próbnymi, czyli $w_i = \phi_i$, co powoduje, że równanie (11.15) przyjmuje postać

$$\int_{\Omega} (\mathrm{D}\tilde{y} - f) \phi_i \mathrm{d}\Omega = 0$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wykorzystując wyrażenia (11.13), powyższe równanie można zapisać w postaci

$$\sum_{i=1}^{n} c_i \int_{\Omega} \mathrm{D}\phi_i \phi_k \mathrm{d}\Omega - \int_{\Omega} f\phi_k \mathrm{d}\Omega = 0, \qquad \text{dla } i = k = 1, 2, \ldots, n \tag{11.16}$$

Układ równań (11.16), nazywany układem Galerkina, można zapisać jako

$$\sum_{i=1}^{n} c_i k_{ik} = f_k, \qquad \text{dla } k = 1, 2, \ldots, n$$

gdzie

$$k_{ik} = \int_{\Omega} \mathrm{D}\phi_i \phi_k \mathrm{d}\Omega \tag{11.17}$$

$$k_{ki} = \int_{\Omega} \mathrm{D}\phi_k \phi_i \mathrm{d}\Omega \tag{11.18}$$

$$f_k = \int_{\Omega} f\phi_k \mathrm{d}\Omega \tag{11.19}$$

$$\tag{11.20}$$

Gdy prawdziwa jest nierówność

$$\int_{\Omega} \mathrm{D}yy \mathrm{d}\Omega \geqslant 0$$

współczynniki $k_{ik} = k_{ki}$, czyli macierz współczynników w układzie równań (11.16) jest symetryczna, co jest bardzo pożądaną właściwością.

Przykład obliczeniowy

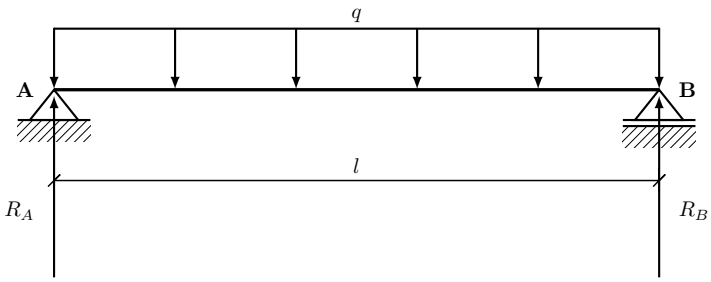
Z powodów wyjaśnionych pod koniec tego przykładu jest on identyczny z przykładem na stronie 215. Dla jasności zamieszczono wszystkie etapy rozwiązania zadania.

Sformułowanie problemu. Metodą Galerkina rozwiązać równanie różniczkowe opisujące linię ugięcia belki pryzmatycznej, jednorodnej, obciążonej jak na rysunku 11.4.

Dane. Znana jest długość belki l, moduł Younga E, moment bezwładności przekrojów belki względem osi prostopadłej do rysunku I oraz wartości sił. Wartości poszczególnych parametrów zestawiono poniżej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wielkość	Wartość	Jednostka
l	1	[m]
q	2000	[N/m]
E	$2{,}1 \cdot 10^9$	[Pa]
$I = bh^3/12$	$8{,}333 \cdot 10^{-6}$	[m^4]



Rysunek 11.4. Schemat obciążenia rozwiązywanej belki

Rozwiązanie

Wyznaczenie reakcji podpór. Aby określić wartości sił reakcji działających na belkę, należy uwolnić ją od więzów, zaś w ich miejsce przyłożyć siły oddziaływania podpór o takich wartościach, aby belka pozostała w stanie równowagi. Równania statyki, które mówią, że suma rzutów wszystkich sił na oś Y ma być zerem i suma momentów sił względem punktu podparcia A ma być zerem, mają postać (11.21).

$$\begin{cases} \sum F_y = R_A - ql + R_B = 0 \\ \sum M_A = -ql\frac{l}{2} + R_B l = 0 \end{cases} \tag{11.21}$$

Rozwiązując układ równań (11.21), można wyznaczyć wartości sił reakcji jako

$$R_A = R_B = \frac{ql}{2}$$

Momenty zginające w przekrojach poprzecznych belki można zatem przedstawić następującą zależnością

$$M(x) = R_A x - \frac{qx^2}{2} = 1000\left(x - x^2\right) \tag{11.22}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Linia ugięcia belki $w(x)$, przy założeniu niewielkich odkształceń, spełnia równanie

$$EI\frac{\mathrm{d}^2w(x)}{\mathrm{d}x^2} + M(x) = 0 \tag{11.23}$$

z warunkami brzegowymi wynikającymi ze sposobu podparcia belki, zgodnie z równaniami (11.24)

$$\begin{aligned} x &= 0 \qquad & w &= 0 \\ x &= l \qquad & w &= 0 \end{aligned} \tag{11.24}$$

Po przekształceniu równanie ugięcia belki można zapisać jako

$$\frac{\mathrm{d}^2w}{\mathrm{d}x^2} = -\frac{R_A - q\frac{x^2}{2}}{EI}$$

Aby zweryfikować dokładność zastosowanej metody przybliżonej, warto wyznaczyć rozwiązanie dokładne tego równania. Po dwukrotnym scałkowaniu i wykorzystaniu warunków brzegowych, linia ugięcia analizowanej belki wyraża się następującym wielomianem stopnia czwartego (11.25).

$$w(x) = \frac{x^4 - 2x^3 + x}{210} \tag{11.25}$$

Oczywiście, jak już wspomniano, te rozważania są prawdziwe przy założeniu niewielkich przemieszczeń.

Rozwiązanie metodą Galerkina

W pierwszym kroku należy wybrać funkcje próbne spełniające warunki brzegowe. Aby zademonstrować istotę metod przybliżonych, jako funkcje próbne zostaną wybrane wielomiany stopni innych niż czwarty.

W pierwszym podejściu zastosowana zostanie jedna funkcja próbna w postaci

$$\phi_1 = x(1 - x^2)$$

zatem przybliżone rozwiązanie jest następujące

$$\tilde{w} = c_1\phi_1 = c_1x(1 - x^2)$$

Po wstawieniu tego przybliżonego rozwiązania do równania (11.23), można wyznaczyć residua jako

$$R(x, c_1) = \frac{\mathrm{d}^2\tilde{w}(x)}{\mathrm{d}x^2} + M(x) = \frac{1000x - 1000x^2}{EI} - 6c_1x \approx \frac{2}{35}(x - x^2) - 6c_1x$$

Następnie, by wyznaczyć wartość parametru c_1, należy wykorzystać warunek zerowania się całki z iloczynu residuum i funkcji wagi

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\int_0^1 \phi_1 R(x,c_1)\mathrm{d}x = \int_0^1 x(1-x^2)\left(\frac{2}{35}(x-x^2) - 6c_1 x\right)\mathrm{d}x =$$

$$= \int_0^1 \frac{2x^5}{35} + 6c_1x^4 - \frac{2x^4}{35} - \frac{2x^3}{35} - 6c_1x^2 + \frac{2x^2}{35}\mathrm{d}x = -\frac{280c_1 - 1}{350} = 0 \quad (11.26)$$

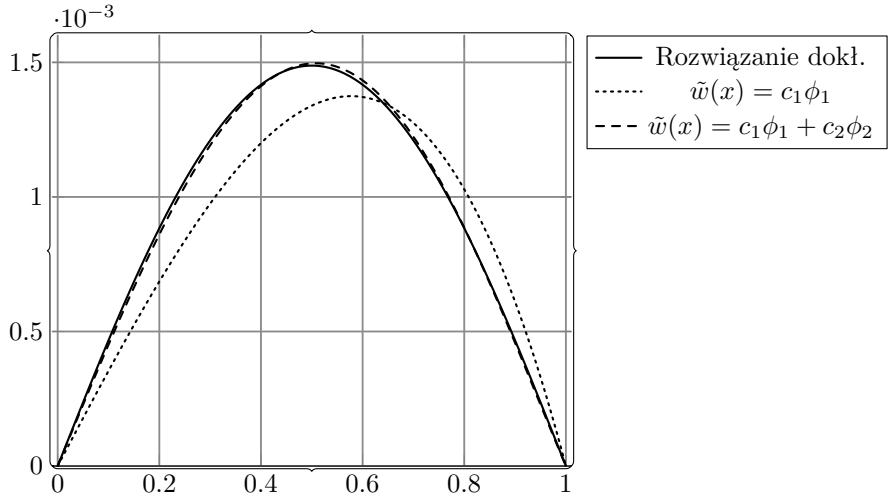
Jak łatwo można stwierdzić, wartość współczynnika c_1 wynosi

$$c_1 = \frac{1}{280}$$

Wobec tego przybliżone rozwiązanie równania różniczkowego ma postać

$$\tilde{w}(x) = \frac{1}{280}x(1-x^2)$$

Porównanie rozwiązania przybliżonego z dokładnym znajduje się na rysunku 11.5.



Rysunek 11.5. Porównanie rozwiązania dokładnego i przybliżonego metodą Galerkina

W drugim podejściu użyte zostaną dwie funkcje próbne. Ponownie wybrane w taki sposób, by zademonstrować przybliżony charakter metody.

$$\begin{cases} \phi_1 = x(1-x^2) \\ \phi_2 = x(1-x^4) \end{cases}$$

Zatem, w tym przypadku, przybliżone rozwiązanie ma postać

$$\tilde{w}(x) = c_1\phi_1 + c_2\phi_2$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Residua wyznaczone po wstawieniu przybliżonego rozwiązania do rozwiązywanego równania różniczkowego przedstawiają się następująco

$$R(x,c_1,c_2)=\frac{\mathrm{d}^2\tilde{w}(x)}{\mathrm{d}x^2}+M(x)=-20c_2x^3+\frac{1000x-1000x^2}{EI}-6c_1x\approx$$
$$\approx -20c_2x^3+\frac{2}{35}(x-x^2)-6c_1x$$

Kolejnym etapem jest ułożenie układu równań, który pozwoli wyznaczyć wartości współczynników stojących przy funkcjach próbnych.

$$\int_0^1 \phi_1 R(x,c_1,c_2)\mathrm{d}x = 0$$
$$\int_0^1 \phi_2 R(x,c_1,c_2)\mathrm{d}x = 0$$

Po wyznaczeniu całek, układ równań wygląda następująco

$$-\frac{-1+280c_1+400c_2}{350} = 0$$
$$-\frac{-33+10\,080c_1+15\,680c_2}{8820} = 0$$

a jego rozwiązaniem jest para liczb

$$c_1=\frac{31}{4480},\qquad c_2=-\frac{3}{1280}$$

Przybliżone rozwiązanie równania linii ugięcia belki przedstawia się w tym przypadku następująco

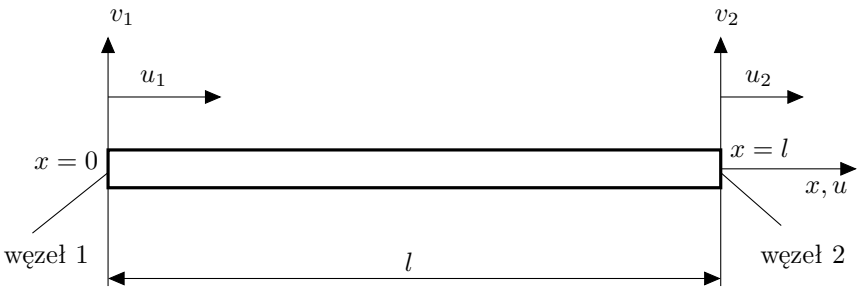
$$\tilde{w}(x)=\frac{31}{4480}x(1-x^2)-\frac{3}{1280}x(1-x^4) \tag{11.27}$$

Rysunek 11.5 przedstawia rozwiązanie dokładne oraz przybliżone dla przypadku rozwiązania z jedną funkcją próbną i z dwoma funkcjami próbnymi.

Warto zwrócić uwagę na osiągniętą w tym przypadku zgodność rozwiązań uzyskanych metodą Rayleigha–Ritza oraz metodą Galerkina. Dla takich samych funkcji próbnych uzyskuje się te same współczynniki, i co za tym idzie, tę samą postać rozwiązania [4].

Po tym krótkim przedstawieniu przybliżonych metod rozwiązywania problemów brzegowych i początkowo-brzegowych opisywanych równaniami różniczkowymi za pomocą metod residuów ważonych i sformułowanych wariacyjnie zostanie przedstawiony sposób wyznaczania macierzy niezbędnych w metodzie elementów skończonych. Na przykładzie pręta i belki przedstawione zostanie wyprowadzenie macierzy sztywności i bezwładności dla pojedynczego elementu. Następnie zostanie przeprowadzona transformacja do globalnego układu współrzędnych i agregacja do macierzy sztywności i mas dla całego układu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.6. Element prętowy

11.5. Wyznaczenie macierzy sztywności elementu prętowego

Pod pojęciem elementu prętowego rozumie się element przenoszący obciążenia wzdłużne, którego jeden wymiar znacząco przewyższa pozostałe.

Wychodząc z zasady minimum energii potencjalnej, można wyznaczyć macierz sztywności elementu prętowego. Aby ją wyznaczyć, przyjmuje się jako obszar rozwiązania element prętowy o długości l przedstawiony na rysunku 11.6. Całkowita energia potencjalna E_p składa się z energii odkształcenia sprężystego U_e i pracy sił zewnętrznych W

$$E_p = U_e + W \tag{11.28}$$

Można wykazać, że dla ciała liniowo-sprężystego E_p jest funkcjonałem kwadratowym i ma jedno globalne minimum [31]. Zgodnie z zasadą minimalizacji energii potencjalnej, otrzymuje się dla elementu następującą zależność

$$\frac{\partial E_p}{\partial \{u_i\}} = 0 \tag{11.29}$$

gdzie $\{u_i\}$ oznacza wektor przemieszczeń węzłów elementu.

Ponieważ pręt przenosi wyłącznie obciążenia osiowe, energia odkształcenia sprężystego wynosi

$$U_e = \int\limits_{\Omega} \frac{\sigma^2}{2E} \mathrm{d}\Omega = \frac{1}{2} \int\limits_{\Omega} \varepsilon^2 E \mathrm{d}\Omega \tag{11.30}$$

gdzie σ jest naprężeniem normalnym, $\mathrm{d}\Omega$ jest elementem objętości, ε jest odkształceniem względnym, E jest modułem Younga. W zapisie macierzowym równanie powyższe przyjmuje postać

$$U_e = \frac{1}{2} \int\limits_{\Omega} \{\varepsilon\}^T [D] \{\varepsilon\} \mathrm{d}\Omega \tag{11.31}$$

gdzie ε jest wektorem odkształceń, zaś $[D]$ jest macierzą współczynników sprężystości. Praca sił zewnętrznych może być zapisana w następującej postaci

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$W = -\{u_i\}^T\{P_i\} \tag{11.32}$$

gdzie $\{P_i\}$ oznacza wektor sił węzłowych.

Z równań (11.31) oraz (11.32) można wyznaczyć wartość całkowitej energii potencjalnej

$$E_p = U_e + W = \frac{1}{2}\int_{\Omega}\{\varepsilon\}^T[D]\{\varepsilon\}\mathrm{d}\Omega - \{u_i\}^T\{P_i\} \tag{11.33}$$

Aby z równania (11.33) wyeliminować odkształcenie ε, można przyjąć, że $[B]$ jest macierzą wiążącą przemieszczenia z odkształceniami, co oznacza, że prawdziwy jest związek

$$\{\varepsilon\} = [B]\{u_i\} \tag{11.34}$$

Zależność (11.34) pozwala na zapisanie równania (11.33) w postaci

$$E_p = \frac{1}{2}\{u_i\}^T\int_{\Omega}[B]^T[D][B]\mathrm{d}\Omega\{u_i\} - \{u_i\}^T\{P_i\} \tag{11.35}$$

Po zróżniczkowaniu i przyrównaniu do zera, zgodnie z równaniem (11.29), otrzymuje się wyrażenie wiążące siłę i przemieszczenia węzłowe

$$\{P_i\} = \int_{\Omega}[B]^T[D][B]\mathrm{d}\Omega\{u_i\} \tag{11.36}$$

Tradycyjnie w mechanice zależność między siłą i odkształceniem jest charakteryzowana przez współczynnik sprężystości $F = kx$. Wobec tego równanie (11.36) w postaci macierzowej można zapisać następująco

$$\{P_i\} = [k]\{u_i\} \tag{11.37}$$

gdzie macierz $[k]$, nazywana macierzą sztywności elementu prętowego, wyraża się wzorem (11.38).

$$[k] = \int_{\Omega}[B]^T[D][B]\mathrm{d}\Omega \tag{11.38}$$

By znaleźć postać macierzy $[B]$, należy wyrazić przemieszczenia poszczególnych przekrojów wzdłuż elementu $\{u\}$ przez przemieszczenia węzłów $\{u_i\}$. Z rysunku 11.6 widać, iż element prętowy ma dwa stopnie swobody: przemieszczenie węzła 1 i przemieszczenie węzła 2. Przyjęto rozkład liniowy przemieszczeń wzdłuż pręta, czyli funkcja interpolująca rozkład przemieszczeń jest wielomianem pierwszego stopnia i ma dwa współczynniki α_1 i α_2

$$u = \alpha_1 + \alpha_2 x \tag{11.39}$$

Aby wyznaczyć wartości współczynników α_1 oraz α_2, należy zadbać, by były zachowane przemieszczenia poszczególnych węzłów u_i elementu, czyli otrzyma się następujące warunki brzegowe

$$\begin{cases} \text{dla } x = 0, \quad u = u_1 \\ \text{dla } x = l, \quad u = u_2 \end{cases} \tag{11.40}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

To jest interpolacja Lagrange'a a zatem element jest nazywa się elementem lagrange'owskim.

Podstawiając warunki dotyczące przemieszczeń na brzegu do równania (11.39), otrzymuje się bezpośrednio

$$\alpha_1 = u_1, \qquad \alpha_2 = \frac{u_2 - u_1}{l}$$

czyli wielomian interpolujący rozkład przemieszczeń wzdłuż pręta ma postać

$$u = u_1 + \frac{u_2 - u_1}{l}x = u_1 + \frac{u_2}{l}x - \frac{u_1}{l}x = u_1\left(1 - \frac{x}{l}\right) + u_2\frac{x}{l} \tag{11.41}$$

Po wprowadzeniu współrzędnej lokalnej jako

$$\xi = \frac{x}{l}$$

można zapisać wielomian interpolujący następująco

$$u = u_1(1 - \xi) + u_2\xi \tag{11.42}$$

Przyjęty wzór wielomianu pozwala na zbudowanie funkcji kształtu spełniającej wszystkie warunki podane na początku tego rozdziału: zapewnia ciągłość przemieszczeń wewnątrz elementu, umożliwia opisanie stałych przemieszczeń oraz stałych odkształceń, zatem opracowywany element będzie elementem zgodnym.

Równanie (11.42) można zapisać macierzowo

$$\{u\} = [N]\{u_i\} \tag{11.43}$$

gdzie $\{u\}$ jest wektorem funkcji przemieszczeń, macierz $[N]$ nazywa się macierzą funkcji kształtu. Macierz funkcji kształtu nosi taką nazwę, ponieważ zawiera funkcje określające rozkład przemieszczeń wzdłuż elementu, może też być używana do określania kształtu elementu. Gdy dla danego elementu macierz funkcji kształtu opisująca przemieszczenia jest identyczna z macierzą funkcji opisujących jego kształt, element nazywa się izoparametrycznym. Macierz funkcji kształtu prętowego elementu liniowego przyjmuje postać

$$[N] = [(1 - \xi) \quad \xi] \tag{11.44}$$

Po określeniu przemieszczeń można wyznaczyć odkształcenia. Wykorzystując definicję odkształcenia, otrzymuje się

$$\varepsilon = \frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}u}{l\mathrm{d}\xi} = \frac{1}{l}\frac{\mathrm{d}[N]}{\mathrm{d}\xi}\{u_i\} \tag{11.45}$$

Po wykorzystaniu równania (11.42) można napisać

$$\varepsilon = \frac{-1}{l}u_1 + \frac{1}{l}u_2 \tag{11.46}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

W postaci macierzowej można w ten oto sposób zapisać tę zależność

$$\{\varepsilon\} = \begin{bmatrix} \frac{-1}{l} & \frac{1}{l} \end{bmatrix} \{u_i\}, \quad \text{gdzie } \{u_i\} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{11.47}$$

Porównując wyrażenie (11.34) z równaniem (11.34), można zauważyć, że

$$[B] = \begin{bmatrix} \frac{-1}{l} & \frac{1}{l} \end{bmatrix}$$

Macierz $[B]$ jest nazywana macierzą zgodności geometrycznej. Z prawa Hooke'a ($\sigma = E\varepsilon$) oraz równania (11.31) wynika następująca zależność na macierz współczynników sprężystości pręta z materiału izotropowego

$$[D] = E$$

Wstawiając powyższe zależności do równania (11.38), uzyskuje się wyrażenie na macierz sztywności w postaci

$$[k] = \int_{\Omega} [B]^T [D][B] \mathrm{d}\Omega \tag{11.48}$$

Pole przekroju poprzecznego pręta wynosi A, wobec tego element objętości pręta można wyrazić następująco $\mathrm{d}\Omega = A\mathrm{d}x = Al\mathrm{d}\xi$. W związku z tym macierz sztywności przyjmuje postać

$$[k] = \int_0^1 \begin{bmatrix} -\frac{1}{l} \\ \frac{1}{l} \end{bmatrix} E \begin{bmatrix} -\frac{1}{l} & \frac{1}{l} \end{bmatrix} lA\mathrm{d}\xi =$$

$$= AEl \int_0^1 \begin{bmatrix} \frac{1}{l^2} & -\frac{1}{l^2} \\ -\frac{1}{l^2} & \frac{1}{l^2} \end{bmatrix} \mathrm{d}\xi = \frac{AEl}{l^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \int_0^1 \mathrm{d}\xi \tag{11.49}$$

Macierz sztywności elementu prętowego leżącego wzdłuż osi OX i poddanego działaniu sił osiowych ma postać

$$[k] = \frac{AE}{l} \begin{array}{c} \begin{matrix} u_1 & u_2 \end{matrix} \\ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{array} \begin{array}{l} u_1 \\ u_2 \end{array} \tag{11.50}$$

Opisy poszczególnych wierszy i kolumn macierzy sztywności dają informację, których węzłów dotyczą poszczególne wiersze i kolumny macierzy. Następnym krokiem będzie wyznaczenie macierzy bezwładności elementu prętowego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

11.6. Wyznaczenie macierzy bezwładności elementu prętowego

Wyznaczanie macierzy bezwładności elementu prętowego dla celów dynamiki rozpoczyna się od określenia energii kinetycznej. Jeżeli m oznacza masę elementu, zaś $\dot{u}$ oznacza prędkość, to energia kinetyczna jest równa

$$E_k = \frac{1}{2} m \dot{u}^2 \tag{11.51}$$

W postaci macierzowej powyższa zależność przyjmuje postać

$$E_k = \frac{1}{2} \{\dot{u}_i\}^T [m] \{\dot{u}_i\} \tag{11.52}$$

gdzie $[m]$ jest macierzą bezwładności elementu, a $\{\dot{u}_i\}$ jest wektorem prędkości węzłowych. Dla całego elementu energia kinetyczna jest całką po całej jego objętości, czyli

$$E_k = \frac{1}{2} \int_\Omega \{\dot{u}\}^T \rho \{\dot{u}\} \mathrm{d}\Omega \tag{11.53}$$

gdzie $\{\dot{u}\}$ jest wektorem zawierającym funkcję zmian prędkości wzdłuż elementu skończonego, natomiast ρ jest gęstością materiału. Niech zostanie przyjęte rozwiązanie w postaci harmonik czasowych, czyli sinusoidalnie zmienne w czasie

$$\{u\} = A e^{i\omega t}$$

gdzie A jest amplitudą, ω częstością kołową, t oznacza czas. Zatem wektor opisujący rozkład prędkości wzdłuż elementu wygląda następująco

$$\{\dot{u}\} = i\omega A e^{i\omega t} = i\omega \{u\}$$

W równaniu (11.43) zapisano zależność na funkcję określającą przemieszczenia elementu jako $\{u\} = [N]\{u_i\}$. Wobec tego

$$\{\dot{u}\} = i\omega [N] \{u_i\}$$

Znając rozkład prędkości ruchu przekrojów wzdłuż pręta z równania (11.53), wyznacza się energię kinetyczną

$$E_k = \frac{1}{2} \int_\Omega i\omega \{u_i\}^T [N]^T \rho i\omega [N] \{u_i\} \mathrm{d}\Omega = -\frac{1}{2} \omega^2 \{u_i\}^T \int_\Omega [N]^T \rho [N] \mathrm{d}\Omega \{u_i\} \tag{11.54}$$

Z porównania równań (11.54) i (11.52) widać, że macierz bezwładności stanowi wyrażenie pod całką

$$[m] = \int_\Omega [N]^T \rho [N] \mathrm{d}\Omega \tag{11.55}$$

Gęstość materiału elementu prętowego nie musi być wartością stałą w całej objętości, jednak gdy jest stała, to może być wyłączona przed znak całki.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Ponieważ funkcje kształtu elementu prętowego są dane równaniem (11.44), można napisać, że

$$[m] = \int_{\Omega} \begin{bmatrix} 1-\xi \\ \xi \end{bmatrix} \rho[(1-\xi) \quad \xi] \mathrm{d}\Omega = \rho A l \int_0^1 \begin{bmatrix} 1-\xi \\ \xi \end{bmatrix} [(1-\xi) \quad \xi] \mathrm{d}\xi =$$

$$= \rho A l \int_0^1 \begin{bmatrix} (1-\xi)^2 & \xi(1-\xi) \\ \xi(1-\xi) & \xi^2 \end{bmatrix} \mathrm{d}\xi = \rho A l \int_0^1 \begin{bmatrix} 1-2\xi+\xi^2 & \xi-\xi^2 \\ \xi-\xi^2 & \xi^2 \end{bmatrix} \mathrm{d}\xi =$$

$$= \rho A l \begin{bmatrix} \xi - \xi^2 + \frac{1}{3}\xi^3 & \frac{1}{2}\xi^2 - \frac{1}{3}\xi^3 \\ \frac{1}{2}\xi^2 - \frac{1}{3}\xi^3 & \frac{1}{3}\xi^3 \end{bmatrix}\Bigg|_0^1 = \rho A l \begin{bmatrix} 1-1+\frac{1}{3} & \frac{1}{2}-\frac{1}{3} \\ \frac{1}{2}-\frac{1}{3} & \frac{1}{3} \end{bmatrix} =$$

$$= \rho A l \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{bmatrix} = \frac{\rho A l}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{11.56}$$

Zatem macierz bezwładności elementu prętowego w ruchu wzdłuż osi x wyraża się zależnością

$$[m] = \frac{\rho A l}{6} \begin{matrix} \\ u_1 \\ u_2 \end{matrix} \begin{matrix} \begin{matrix} u_1 & u_2 \end{matrix} \\ \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \end{matrix} \tag{11.57}$$

Wokół macierzy bezwładności podano stopnie swobody, z którymi są powiązane odpowiednie wiersze i kolumny macierzy. Następnym etapem obliczeń jest agregacja elementarnych macierzy sztywności i bezwładności do macierzy globalnych.

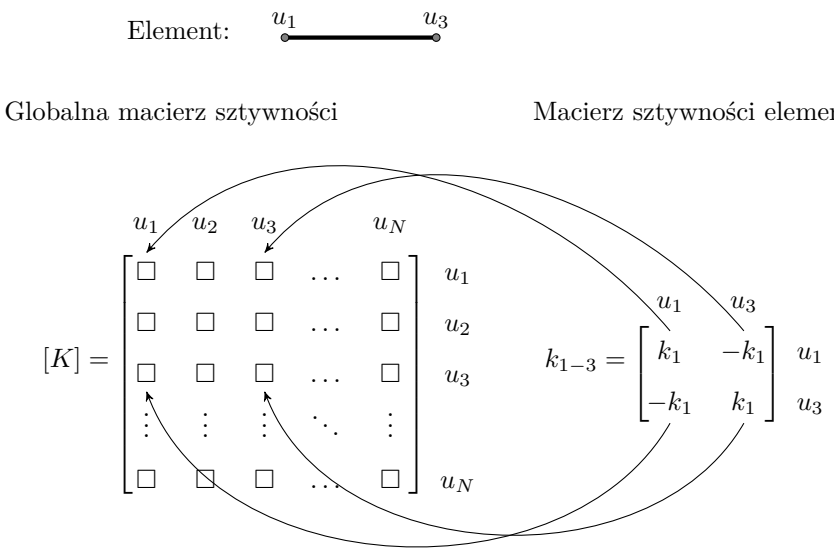
11.7. Agregacja macierzy sztywności i bezwładności

Zazwyczaj analizowany układ składa się ze znacznie większej liczby elementów niż jeden. W związku z tym układ ma znacznie więcej stopni swobody, czyli szukanych wielkości, które wyznacza się w węzłach. Dlatego macierze sztywności i bezwładności dla całego układu będą znacznie większe niż te, które do tej pory przedstawiono. Ich rozmiary będą równe iloczynowi liczby stopni swobody, czyli zmiennych w węźle, oraz liczby węzłów.

Tworzenie macierzy opisującej cały układ, zwanej macierzą globalną nazywa się agregacją. Agregację należy przeprowadzić dla wszystkich macierzy opisujących elementy, czyli dla macierzy sztywności, bezwładności i ewentualnie macierzy tłumienia. Zanim będzie można dokonać agregacji, należy macierze elementów wyrazić w globalnym układzie odniesienia. W analizowanym przypadku globalny układ odniesienia pokrywa się z lokalnym.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Globalna macierz sztywności czy bezwładności zawiera wiersze i kolumny opisujące wszystkie stopnie swobody układu. Agregacja polega na tym, by do macierzy globalnych wpisać w odpowiednie miejsca elementy lokalnych macierzy dotyczących poszczególnych elementów. Schematycznie agregację macierzy sztywności przedstawiono to na rysunku 11.7.



Rysunek 11.7. Agregacja macierzy sztywności dla elementu liniowego z węzłami o numerach 1 i 3

Dla elementu, którego węzły mają numery 1 oraz 3 wyznaczono macierz sztywności elementu oznaczoną jako k_{1-3}. Elementy tej macierzy są przypisane odpowiednim węzłom. Strzałki na rysunku 11.7 ilustrują sposób wpisywania poszczególnych elementów macierzy sztywności elementu do globalnej macierzy sztywności.

W zagadnieniach statycznych macierz sztywności stanowi macierz współczynników układu równań, w którym niewiadomymi są przemieszczenia węzłów, zaś prawa strona zawiera zewnętrzne siły przykładane do węzłów. W ten sposób buduje się układ liniowych równań algebraicznych opisujących stan równowagi układu, który następnie trzeba rozwiązać, by wyznaczyć przemieszczenia w węzłach.

11.7.1. Nakładanie warunków brzegowych Dirichleta

Warunki brzegowe Dirichleta, oznaczające wartość zmiennej w węźle, nakłada się modyfikując otrzymany układ równań. Modyfikacje dotyczą macierzy sztywności i macierzy obciążeń węzłowych. Jeżeli w i-tym węźle należy przypisać przemieszczeniu wartość U, to modyfikacja układu równań będzie przebiegała jak w równaniu (11.58).

$$u_i = U \tag{11.58}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = U \tag{11.59}$$

Równanie (11.58) przedstawia przemieszczenie u_i zapisane bez uwzględnienia innych stopni swobody natomiast równanie (11.59) – z ich uwzględnieniem. To równanie należy uzupełnić równaniami dotyczącymi pozostałych węzłów. Taki pełny układ równań opisujący przemieszczenia całego układu, ze wszystkimi stopniami swobody i wszystkimi siłami węzłowymi wygląda tak, jak pokazano poniżej w równaniu (11.60).

$$\begin{bmatrix} a_{11} & \cdots & a_{1i-1} & a_{1i} & a_{1i+1} & \cdots & a_{1n} \\ \vdots & & & \vdots & & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & & \vdots \\ a_{n1} & \cdots & a_{ni-1} & a_{ni} & a_{ni+1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} P_1 \\ \vdots \\ U \\ \vdots \\ P_n \end{bmatrix} \tag{11.60}$$

Otrzymana w ten sposób globalna macierz sztywności nie jest już symetryczna. Z uwagi na możliwość wykorzystania szybkich metod rozwiązywania układów równań z macierzami symetrycznymi warto dokonać zabiegu pozwalającego na przywrócenie symetrii macierzy. Na przykład wpisać zera w kolumnie odpowiadającej modyfikowanemu stopniowi swobody. Koniecznie trzeba też uwzględnić tę zmianę w prawej stronie równania, dokonując odpowiedniej modyfikacji wektora sił węzłowych.

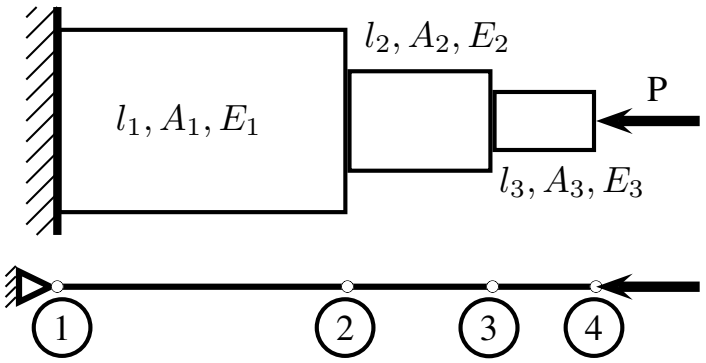
$$\begin{bmatrix} a_{11} & \cdots & a_{1i-1} & 0 & a_{1i+1} & \cdots & a_{1n} \\ \vdots & & & \vdots & & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots & & & \vdots \\ a_{n1} & \cdots & a_{ni-1} & 0 & a_{ni+1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} P_1 - Ua_{1i} \\ \vdots \\ U \\ \vdots \\ P_n - Ua_{ni} \end{bmatrix} \tag{11.61}$$

Oczywiste się staje, że gdy danemu stopniowi swobody przypisuje się przemieszczenie równe zeru, to prawa strona równania jest jedynie zerowana w wierszu odpowiadającym odbieranemu stopniowi swobody.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

11.7.2. Przykład rozwiązania jednowymiarowego problemu statycznego

Sformułowanie problemu. Znaleźć zmiany długości poszczególnych bloków tworzących strukturę jak na rysunku 11.8. Nie uwzględniać wyboczenia. Siła obciążająca wynosi $P = 10^4$ N.



Rysunek 11.8. Analizowana struktura i model elementów skończonych

Rozwiązanie. Problem jest jednowymiarowy, można zatem zastosować jednowymiarowe elementy prętowe. Model elementów skończonych będzie się składał z trzech elementów i będzie wymagał wyznaczenia przemieszczeń w czterech węzłach. Na rysunku 11.8 oprócz schematu zadania, przedstawiono też model elementów skończonych. Każdy węzeł ma swój numer. Macierze elementów są opisywane numerami węzłów, które te elementy łączą. Czyli element, którego końce stanowią węzły 2 oraz 3 ma macierz sztywności k_{2-3}. Charakterystykę elementów zestawiono poniżej.

	Współrzędne				
i	W1	W2	l_i [m]	E_i [Pa]	A_i [m^2]
1	0	0,40	0,40	$2 \cdot 10^{11}$	$3{,}6 \cdot 10^{-3}$
2	0,40	0,65	0,25	$2 \cdot 10^{9}$	$2{,}5 \cdot 10^{-3}$
3	0,65	0,85	0,20	$2 \cdot 10^{7}$	$1{,}6 \cdot 10^{-3}$

Problem jest statyczny, więc równanie opisujące równowagę układu ma postać

$$[k]\{u_i\} = \{P_i\} \tag{11.62}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie wektor przemieszczeń węzłowych składa się z czterech elementów

$$\{u_i\} = \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{Bmatrix}$$

Macierze sztywności dla poszczególnych elementów skończonych zostały przedstawione poniżej. Poszczególne wiersze i kolumny macierzy zostały opisane stopniami swobody, do których się odnoszą.

$$k_{1-2} = \frac{A_1 E_1}{l_1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{matrix} \\ u_1 \\ u_2 \end{matrix} \begin{matrix} \begin{matrix} u_1 & u_2 \end{matrix} \\ \begin{bmatrix} k_1 & -k_1 \\ -k_1 & k_1 \end{bmatrix} \end{matrix}$$

W macierzy k_{1-2} pierwsza kolumna dotyczy przesunięcia węzła 1, druga kolumna przesunięciu węzła 2, pierwszy wiersz dotyczy równania dla węzła 1, drugi wiersz za równanie dla węzła 2.

$$k_{2-3} = \frac{A_2 E_2}{l_2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{matrix} \\ u_2 \\ u_3 \end{matrix} \begin{matrix} \begin{matrix} u_2 & u_3 \end{matrix} \\ \begin{bmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} \end{matrix}$$

Analogicznie do poprzedniej macierzy, pierwsza kolumna dotyczy przesunięcia węzła 2, druga przesunięcia węzła 3, pierwszy wiersz to współczynniki równania równowagi dla węzła 2, drugi dla węzła 3,

$$k_{3-4} = \frac{A_3 E_3}{l_3} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{matrix} \\ u_3 \\ u_4 \end{matrix} \begin{matrix} \begin{matrix} u_3 & u_4 \end{matrix} \\ \begin{bmatrix} k_3 & -k_3 \\ -k_3 & k_3 \end{bmatrix} \end{matrix}$$

gdzie $k_1 = 1{,}8 \cdot 10^9$, $k_2 = 2{,}0 \cdot 10^7$ oraz $k_3 = 1{,}6 \cdot 10^5$. Poszczególne wyrazy macierzy sztywności poszczególnych elementów należy wpisywać w odpowiednie pola macierzy

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

globalnej zgodnie z tym, których węzłów dotyczą. Zatem macierz sztywności dla całego układu przedstawia się następująco

$$[k] = \begin{matrix} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{matrix} \begin{matrix} \begin{matrix} u_1 & u_2 & u_3 & u_4 \end{matrix} \\ \begin{bmatrix} k_1 & -k_1 & 0 & 0 \\ -k_1 & k_1+k_2 & -k_2 & 0 \\ 0 & -k_2 & k_2+k_3 & -k_3 \\ 0 & 0 & -k_3 & k_3 \end{bmatrix} \end{matrix}$$

Utwierdzenie układu należy uwzględnić w postaci warunku brzegowego Dirichleta. Nałożenie warunków brzegowych Dirichleta, czyli przemieszczenie pierwszego węzła równe zero, uzyskuje się, wstawiając wartość 1 w odpowiednie miejsce na głównej przekątnej i zerując pozostałe elementy wiersza. Ponieważ ta operacja zaburza symetrię macierzy, więc aby ją zachować, należy podobnie wyzerować elementy w kolumnie, w której wstawiono wartość 1. Aby nie zmienić wartości rozwiązania układu równań, należy od prawych stron odjąć wartość przypisaną danemu stopniowi swobody. Ponieważ stopień swobody jest odbierany, czyli przyjmuje się dla niego przemieszczenie równe zeru, nie jest konieczne modyfikowanie wektora prawych stron układu równań.

Wobec tego zeruje się pierwszy wiersz, podobnie zeruje się pierwszą kolumnę, zaś na przecięciu pierwszego wiersza i pierwszej kolumny wstawia się jedynkę. Końcowa postać macierzy $[k]$ wygląda następująco

$$[k] = \begin{matrix} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{matrix} \begin{matrix} \begin{matrix} u_1 & u_2 & u_3 & u_4 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & k_1+k_2 & -k_2 & 0 \\ 0 & -k_2 & k_2+k_3 & -k_3 \\ 0 & 0 & -k_3 & k_3 \end{bmatrix} \end{matrix}$$

Siła zewnętrzna jest przyłożona do węzła numer 4, zatem wektor obciążeń węzłowych ma postać

$$\{P_i\} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -10^4 \end{bmatrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Na tym etapie można przystąpić do rozwiązania równania (11.62) opisującego przemieszczenia węzłów układu.

$$\{u_i\} = [k]^{-1}\{P_i\}$$

Wyznaczone przemieszczenia poszczególnych węzłów wynoszą odpowiednio

$$\{u_i\} = 10^{-6} \cdot \begin{bmatrix} 0{,}0 \\ -5{,}6 \\ -505{,}6 \\ -63005{,}6 \end{bmatrix}$$

Pobieżna analiza wyników wskazuje, że przemieszczenia poszczególnych węzłów intuicyjnie są zgodne z wartościami stałych materiałowych przypisanych poszczególnym elementom. Trzeba zdawać sobie sprawę, że analizowany model elementów skończonych nie uwzględnia zmian naprężeń wywołanych sposobem obciążenia poszczególnych elementów.

11.7.3. Macierz sztywności i bezwładności w globalnym układzie odniesienia

Rozważa się punkt A o współrzędnych (x^o, y^o) w lokalnym układzie odniesienia obróconym o kąt θ w stosunku do globalnego układu odniesienia. Jaki jest związek między współrzędnymi (x_g^o, y_g^o) tego punktu w globalnym układzie odniesienia oraz współrzędnymi (x^o, y^o) w lokalnym układzie odniesienia? Aby uzyskać odpowiedź na to pytanie, zakłada się, iż punkt A leży w pierwszej ćwiartce obu układów współrzędnych. Analiza rysunku 11.9 wskazuje, iż odpowiednie zależności wyglądają następująco

$$\begin{cases} x^o &= x_g^o \cos\theta + y_g^o \sin\theta \\ y^o &= -x_g^o \sin\theta + y_g^o \cos\theta \end{cases} \tag{11.63}$$

Równanie (11.63) w zapisie macierzowym przyjmuje postać

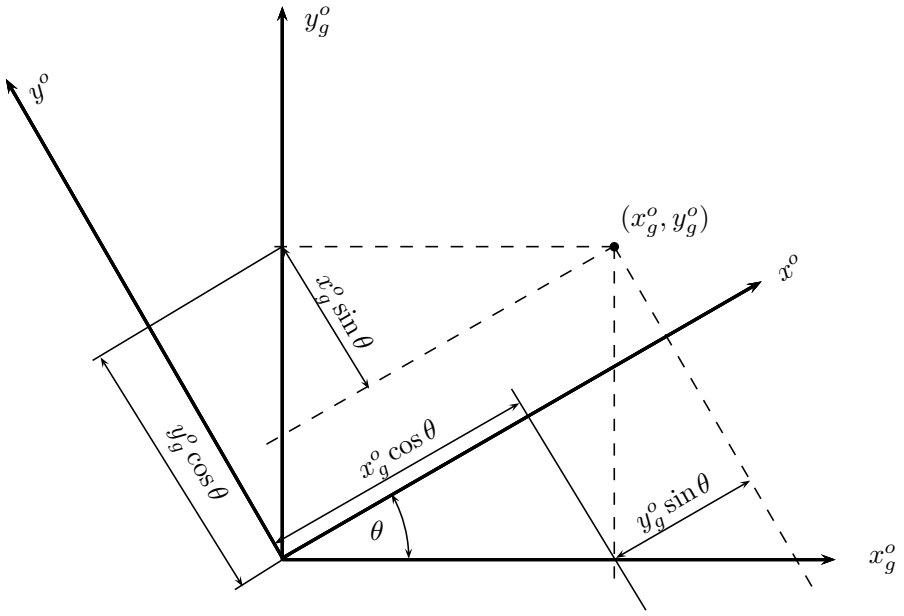
$$\begin{Bmatrix} x^o \\ y^o \end{Bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{Bmatrix} x_g^o \\ y_g^o \end{Bmatrix} \tag{11.64}$$

Gdy współrzędne w układzie globalnym oznaczy się u_i^o, zaś współrzędne w układzie lokalnym u_i, można zapisać macierzową zależność

$$\{u_i\} = [\zeta]\{u_i^o\} \tag{11.65}$$

gdzie ζ oznacza macierz obrotu dla jednego punktu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.9. Współrzędne w globalnym i lokalnym układzie odniesienia

Ponieważ element prętowy ma dwa węzły, z których każdy jest opisany dwoma współrzędnymi, wobec tego przekształcenie przyjmuje postać

$$\begin{Bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{Bmatrix} = \left[\begin{array}{cc|cc} \cos\theta & \sin\theta & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ -\sin\theta & \cos\theta & & \\ \hline \multicolumn{2}{c|}{\multirow{2}{*}{0}} & \cos\theta & \sin\theta \\ & & -\sin\theta & \cos\theta \end{array}\right] \begin{Bmatrix} u_1^o \\ v_1^o \\ u_2^o \\ v_2^o \end{Bmatrix} \tag{11.66}$$

lub w zapisie macierzowym

$$\{u_i\} = [DC]\{u_i^o\}$$

gdzie $[DC]$ – macierz kosinusów kierunkowych równa

$$[DC] = \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{bmatrix} \tag{11.67}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Jest to macierz ortogonalna, co oznacza, że jej macierz odwrotna jest równa transponowanej

$$[DC]^{-1} = [DC]^T$$

Ponieważ znana jest macierz obrotu układu współrzędnych, można napisać, że macierz obciążeń węzłowych w lokalnym układzie odniesienia jest równa

$$\{P_i\} = [DC]\{P_i^o\}$$

gdzie $\{P_i\}$ jest macierzą obciążeń węzłowych w lokalnym układzie współrzędnych, zaś $\{P_i^o\}$ jest macierzą obciążeń węzłowych w globalnym układzie współrzędnych. Skoro w lokalnym układzie współrzędnych prawdziwa jest zależność

$$\{P_i\} = [k]\{u_i\}$$

to używając macierzy obciążeń i macierzy sztywności w globalnym układzie odniesienia, można ją wyrazić w następujący sposób

$$[DC]\{P_i^o\} = [k][DC]\{u_i^o\} \tag{11.68}$$

Otrzymuje się zależność

$$\{P_i^o\} = [DC]^{-1}[k][DC]\{u_i^o\} = [DC]^T[k][DC]\{u_i^o\} \tag{11.69}$$

Powyższe równanie (11.69) można zapisać

$$\{P_i^o\} = [k^o]\{u_i^o\}$$

gdzie

$$[k^o] = [DC]^T[k][DC] \tag{11.70}$$

jest macierzą sztywności w globalnym układzie odniesienia. Podobną zależność można wyprowadzić dla macierzy bezwładności wyrażonej w globalnym układzie odniesienia

$$[m^o] = [DC]^T[m][DC] \tag{11.71}$$

Macierz sztywności dwuwymiarowego elementu prętowego wygląda następująco

$$[k] = \frac{AE}{l}\begin{array}{c} \begin{matrix} u_1 & v_1 & u_2 & v_2 \end{matrix} \\ \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{matrix}$$

Macierz kosinusów kierunkowych zapisano w równaniu (11.67), wobec tego otrzyma się następujące wyrażenie na macierz sztywności w układzie globalnym

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$[k^o] = [DC]^T[k][DC] = \frac{AE}{l}\begin{bmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & -\sin\theta \\ 0 & 0 & \sin\theta & \cos\theta \end{bmatrix} \cdot$$

$$\cdot \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & \cos\theta & \sin\theta \\ 0 & 0 & -\sin\theta & \cos\theta \end{bmatrix} =$$

$$= \frac{AE}{l}\begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta & -\cos^2\theta & -\sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta & -\sin\theta\cos\theta & -\sin^2\theta \\ -\cos^2\theta & -\sin\theta\cos\theta & \cos^2\theta & \sin\theta\cos\theta \\ -\sin\theta\cos\theta & -\sin^2\theta & \sin\theta\cos\theta & \sin^2\theta \end{bmatrix} \tag{11.72}$$

Odpowiednie funkcje trygonometryczne można wyrazić za pomocą współrzędnych węzłów jako

$$\begin{aligned} l &= \sqrt{(x_2^o - x_1^o)^2 + (y_2^o - y_1^o)^2} \\ \cos\theta &= \tfrac{x_2^o - x_1^o}{l} \\ \sin\theta &= \tfrac{y_2^o - y_1^o}{l} \end{aligned} \tag{11.73}$$

Aby wyznaczyć macierz bezwładności w globalnym układzie odniesienia, należy wyznaczyć jej postać przy przemieszczeniach węzłów zarówno w poziomie, jak i w pionie, co prowadzi do następującej postaci funkcji kształtu

$$[N] = \begin{bmatrix} 1-\xi & 0 & \xi & 0 \\ 0 & 1-\xi & 0 & \xi \end{bmatrix}$$

W związku z tym przemieszczenia poszczególnych przekrojów wzdłuż elementu opisać można zależnością

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1-\xi & 0 & \xi & 0 \\ 0 & 1-\xi & 0 & \xi \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{bmatrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zatem macierz bezwładności, na podstawie zależności (11.55), przedstawia się następująco

$$[m] = \int_0^1 N^T \rho N A l d\xi = \rho A l \int_0^1 \begin{bmatrix} 1-\xi & 0 \\ 0 & 1-\xi \\ \xi & 0 \\ 0 & \xi \end{bmatrix} \begin{bmatrix} 1-\xi & 0 & \xi & 0 \\ 0 & 1-\xi & 0 & \xi \end{bmatrix} \mathrm{d}\xi =$$

$$= \rho A l \int_0^1 \begin{bmatrix} (\xi-1)^2 & 0 & -\xi(\xi-1) & 0 \\ 0 & (\xi-1)^2 & 0 & -\xi(\xi-1) \\ -\xi(\xi-1) & 0 & \xi^2 & 0 \\ 0 & -\xi(\xi-1) & 0 & \xi^2 \end{bmatrix} \mathrm{d}\xi =$$

$$= \rho A l \begin{bmatrix} \frac{(\xi-1)^3}{3} & 0 & -\frac{\xi^2(2\xi-3)}{6} & 0 \\ 0 & \frac{(\xi-1)^3}{3} & 0 & -\frac{\xi^2(2\xi-3)}{6} \\ -\frac{\xi^2(2\xi-3)}{6} & 0 & \frac{\xi^3}{3} & 0 \\ 0 & -\frac{\xi^2(2\xi-3)}{6} & 0 & \frac{\xi^3}{3} \end{bmatrix}_0^1 = \frac{\rho A l}{6} \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix} \tag{11.74}$$

Wykorzystując zależność (11.71) oraz macierz obrotu, można wyznaczyć macierz bezwładności w układzie globalnym. Po wymnożeniu otrzyma się

$$[m] = \frac{\rho A l}{6} \begin{array}{c} \begin{matrix} u_1 & v_1 & u_2 & v_2 \end{matrix} \\ \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{matrix} \tag{11.75}$$

Porównanie wyrażeń na macierz bezwładności w układzie odniesienia lokalnym oraz globalnym pozwala na stwierdzenie, że dla elementu prętowego nie ulega ona zmianie przy przejściu z lokalnego do globalnego układu współrzędnych.

Nabytą do tej pory wiedzę o MES można wykorzystać do rozwiązywania zagadnień statycznych, co zostanie przedstawione w podrozdziale 11.9.

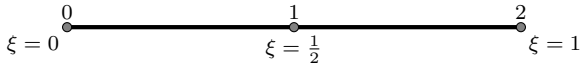
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

11.8. Elementy wyższych rzędów

Oprócz elementów liniowych (dwuwęzłowych), w których zakłada się liniową zmianę wartości wielkości analizowanej wzdłuż krawędzi elementu, można stosować elementy wyższych rzędów. Poniżej zostanie przedstawiony przykład dla elementu parabolicznego oraz sześciennego.

Postępując podobnie jak poprzednio, można wyprowadzić równania elementu dla większej liczby węzłów w elemencie prętowym w lokalnym układzie odniesienia. Dla uogólnienia wyprowadzeń zostanie wykorzystana interpolacja Lagrange'a wyrażona wzorem (2.13).

Zostanie przyjęty lokalny układ współrzędnych ze zmienną $\xi \in [0,1]$ i węzłami na końcach elementu oraz na jego środku. Element przedstawiono schematycznie na rysunku 11.10.



Rysunek 11.10. Element prętowy z paraboliczną funkcją kształtu

Węzły mają współrzędne $\xi_0 = 0$, $\xi_1 = 0{,}5$, $\xi_2 = 1$. W takiej sytuacji można napisać następujące funkcje kształtu, wynikające wprost z interpolacji Lagrange'a

$$N = [l_0 \quad l_1 \quad l_2] = \left[\frac{\xi - \xi_1}{\xi_0 - \xi_1}\frac{\xi - \xi_2}{\xi_0 - \xi_2} \quad \frac{\xi - \xi_0}{\xi_1 - \xi_0}\frac{\xi - \xi_2}{\xi_1 - \xi_2} \quad \frac{\xi - \xi_0}{\xi_2 - \xi_0}\frac{\xi - \xi_1}{\xi_2 - \xi_1} \right] = \\ = [1 - 3\xi + 2\xi^2 \quad 4(\xi - \xi^2) \quad 2\xi(\xi - 1)] \quad (11.76)$$

Przy takiej postaci funkcji kształtu można wyznaczyć macierz sztywności, pamiętając, że macierz zgodności geometrycznej

$$[B] = \frac{1}{l}\frac{\mathrm{d}N}{\mathrm{d}\xi}$$

czyli

$$[B] = \frac{1}{l}\begin{bmatrix} -3 + 4\xi & 4 - 8\xi & 4\xi - 1 \end{bmatrix}$$

co daje macierz sztywności w postaci

$$[k] = \int_{\Omega} [B]^T E [B] \mathrm{d}\Omega = \frac{EA}{3l}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \quad (11.77)$$

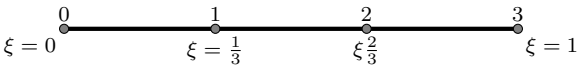
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Podobnie macierz bezwładności można wyliczyć jako

$$[m] = \int_{\Omega} N^T \rho N \mathrm{d}\Omega = \frac{\rho A l}{30} \begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} \tag{11.78}$$

W ten sposób można wyznaczać macierze sztywności oraz bezwładności dla elementów wyższych rzędów. Warto pamiętać, iż wyższy rząd elementu daje znaczne korzyści obliczeniowe.

Element prętowy sześcienny pozwala na modelowanie odkształceń wielomianami trzeciego stopnia. Aby zbudować taką interpolację, element musi mieć cztery węzły jak to pokazano na rysunku 11.11. Dla każdego węzła należy zapisać funkcję kształtu przyjmującą wartość równą jeden w tym węźle i zerującą się w innych.



Rysunek 11.11. Element prętowy z sześcienną funkcją kształtu

Funkcje kształtu, uzyskane za pomocą wielomianów Lagrange'a, są następujące

$$N^T = \begin{bmatrix} -\frac{9}{2}\xi^3 + 9\xi^2 - \frac{11}{2}\xi + 1 \\ \frac{27}{2}\xi^3 - \frac{45}{2}\xi^2 + 9\xi \\ -\frac{27}{2}\xi^3 + 18\xi^2 - \frac{9}{2}\xi \\ \frac{9}{2}\xi^3 - \frac{9}{2}\xi^2 + \xi \end{bmatrix} \tag{11.79}$$

Natomiast macierze sztywności i bezwładności w układzie lokalnym przyjmują następującą postać

$$[k] = \frac{AE}{l} \begin{bmatrix} \frac{37}{10} & \frac{-189}{40} & \frac{27}{20} & \frac{-13}{40} \\ \frac{-189}{40} & \frac{54}{5} & \frac{-297}{40} & \frac{27}{20} \\ \frac{27}{20} & \frac{-297}{40} & \frac{54}{5} & \frac{-189}{40} \\ \frac{-13}{40} & \frac{27}{20} & \frac{-189}{40} & \frac{37}{10} \end{bmatrix} \tag{11.80}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$[m] = \rho A l \begin{bmatrix} \frac{8}{105} & \frac{33}{560} & \frac{-3}{140} & \frac{7}{619} \\ \frac{33}{560} & \frac{27}{70} & \frac{-27}{560} & \frac{-3}{140} \\ \frac{-3}{140} & \frac{-27}{560} & \frac{27}{70} & \frac{33}{560} \\ \frac{7}{619} & \frac{-3}{140} & \frac{33}{560} & \frac{8}{105} \end{bmatrix} \tag{11.81}$$

Elementy wyższych rzędów, mimo znacznie większego nakładu obliczeniowego związanego z większą liczbą węzłów, pozwalają na znacznie efektywniejsze rozwiązywanie złożonych problemów. Wynika to z faktu, że funkcje kształtu w postaci wielomianów wyższych stopni mogą lepiej odwzorować rzeczywiste rozkłady pól przemieszczeń.

11.9. Rozwiązywanie prostych zagadnień statyki

Zagadnienia statyczne sprowadzają się do rozwiązania równań, w których nie występują pochodne badanych wielkości po czasie. Zazwyczaj równanie ma postać

$$[k]\{u_i\} = \{P_i\}$$

gdzie $[k]$ jest macierzą sztywności, $\{u_i\}$ jest poszukiwaną macierzą przemieszczeń węzłowych, zaś $\{P_i\}$ jest znaną macierzą obciążeń. Jak pokazano, macierz sztywności ma postać zależną od typu elementu, a dokładniej od interpolacji przemieszczeń wewnątrz elementu.

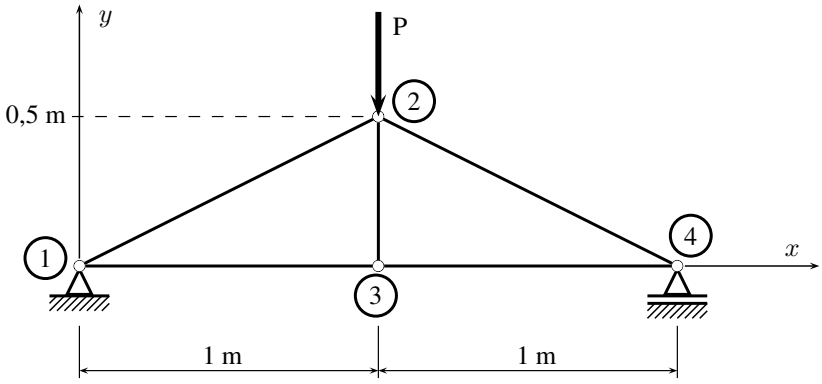
11.9.1. Przykład statyczny płaski

Sformułowanie problemu. Znaleźć wartości przemieszczeń poszczególnych węzłów kratownicy przedstawionej na rysunku 11.12. Wartości współczynników materiałowych przyjąć jak dla stali, czyli $E = 2 \cdot 10^{11}$ Pa. Przekrój poprzeczny wszystkich prętów jest jednakowy i wynosi $A = 8 \cdot 10^{-6}$ m^2. Przemieszczenia węzłów przedstawić w formie graficznej.

Dane. W tabeli 11.1 zebrano dane podane w zadaniu. Długości elementów i wartości kątów wyznaczono na podstawie wzorów (11.73).

Wyznaczenie macierzy sztywności elementów w układzie globalnym będzie realizowane na podstawie zależności (11.72), dla każdego elementu z osobna. Odpowiednie wartości funkcji trygonometrycznych przyjmuje się z tabeli 11.1. Przemieszczenia wzdłuż osi x oznaczono przez u, a te wzdłuż osi y oznaczone zostały przez v. Indeksy przy odpowiednich przemieszczeniach informują o numerze węzła, którego dotyczy dane przemieszczenie. Zatem u_1 oznacza przemieszczenie węzła numer 1 wzdłuż osi x, zaś v_3 oznacza pionowe przemieszczenie węzła numer 3.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.12. Ilustracja do przykładu obliczeniowego

Tabela 11.1. Parametry materiałowe i geometryczne elementów kratownicy z rysunku 11.12

Element	Współrzędne		Długość	Wartość		E [Pa]	A [m^2]
	węzła 1	węzła 2	elementu	$\cos\theta$	$\sin\theta$		
1-2	$(0,0)$	$(1,0{,}5)$	1,118	0,894	0,447	$2\cdot10^{11}$	$8\cdot10^{-6}$
1-3	$(0,0)$	$(1,0)$	1,0	1,0	0,0	$2\cdot10^{11}$	$8\cdot10^{-6}$
2-3	$(1,0{,}5)$	$(1,0)$	0,5	0,0	$-1{,}0$	$2\cdot10^{11}$	$8\cdot10^{-6}$
2-4	$(1,0{,}5)$	$(2,0)$	1,118	0,894	$-0{,}447$	$2\cdot10^{11}$	$8\cdot10^{-6}$
3-4	$(1,0)$	$(2,0)$	1,0	1,0	0,0	$2\cdot10^{11}$	$8\cdot10^{-6}$

Macierz sztywności $[k^o_{1-2}]$ elementu, którego końce stanowią węzły 1 oraz 2, ma niezerowe elementy tylko na przecięciu kolumny 1 i 2 oraz wiersza 1 i 2. Zatem lewy górny fragment macierzy $[k^o_{1-2}]$ może być niezerowy. Pozostałe elementy macierzy są zerami.

Na przykładzie macierzy k^o_{1-3} elementu, którego końcami są węzły 1 oraz 3, przedstawiono sposób wpisywania macierzy sztywności w rozmiarze 4×4 do macierzy 8×8. Omawianą macierz sztywności przedstawia formuła (11.82).

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$[k^o_{1-2}] = 10^6 \cdot \begin{array}{c} \begin{matrix} u_1 & v_1 & u_2 & v_2 & u_3 & v_3 & u_4 & v_4 \end{matrix} \\ \begin{bmatrix} 11{,}2 & 5{,}58 & -11{,}2 & -5{,}58 & 0 & 0 & 0 & 0 \\ 5{,}58 & 2{,}79 & -5{,}58 & -2{,}79 & 0 & 0 & 0 & 0 \\ -11{,}2 & -5{,}58 & 11{,}2 & 5{,}58 & 0 & 0 & 0 & 0 \\ -5{,}58 & -2{,}79 & 5{,}58 & 2{,}79 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ u_4 \\ v_4 \end{matrix}$$

$$[k^o_{1-3}] = 10^7 \cdot \begin{array}{c} \begin{matrix} u_1 & v_1 & u_3 & v_3 \end{matrix} \\ \begin{bmatrix} 1{,}56 & 0 & -1{,}56 & 0 \\ 0 & 0 & 0 & 0 \\ -1{,}56 & 0 & 1{,}56 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ u_3 \\ v_3 \end{matrix}$$

$$[k^o_{1-3}] = 10^7 \cdot \begin{array}{c} \begin{matrix} u_1 & v_1 & u_2 & v_2 & u_3 & v_3 & u_4 & v_4 \end{matrix} \\ \begin{bmatrix} 1{,}56 & 0 & 0 & 0 & -1{,}56 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1{,}56 & 0 & 0 & 0 & 1{,}56 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \\ u_4 \\ v_4 \end{matrix} \tag{11.82}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$[k^o_{2-3}] = 10^7 \cdot \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3{,}12 & 0 & -3{,}12 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3{,}12 & 0 & 3{,}12 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[k^o_{2-4}] = 10^6 \cdot \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 11{,}2 & -5{,}58 & 0 & 0 & -11{,}2 & 5{,}58 \\ 0 & 0 & -5{,}58 & 2{,}79 & 0 & 0 & 5{,}58 & -2{,}79 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -11{,}2 & 5{,}58 & 0 & 0 & 11{,}2 & -5{,}58 \\ 0 & 0 & 5{,}58 & -2{,}79 & 0 & 0 & -5{,}58 & 2{,}79 \end{bmatrix}$$

$$[k^o_{3-4}] = 10^7 \cdot \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1{,}56 & 0 & -1{,}56 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1{,}56 & 0 & 1{,}56 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

W zależności (11.82) ramkami obwiedziono fragmenty macierzy wyliczonej za pomocą wzoru (11.70), a strzałki ilustrują miejsca, w których należy wpisać odpowiednie

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

elementy. Aby nie zaciemnić rysunku, zaznaczono tylko dwa obszary, pozostałe wpisuje się analogicznie.

Po wyznaczeniu macierzy sztywności dla elementów w układzie globalnym, można je zapisać w postaci pozwalającej na wygodną agregację. Jak pokazano wcześniej, macierze o rozmiarach 4×4 dla każdego elementu z osobna należy zapisać w macierzach o rozmiarach 8×8. W dużych układach taki sposób agregacji jest nieekonomiczny. Użyto go w tym przykładzie jedynie ze względów dydaktycznych.

Porównując przedstawione macierze z rysunkiem kratownicy 11.12, można zauważyć, że topologia kratownicy znajduje swoje odwzorowanie w macierzy sztywności. Z macierzy można odczytać, które węzły łączy element. Można też stwierdzić, czy dany element leży wzdłuż osi x, jak element $1-3$, czy też jest ukośny, jak element $1-2$.

Agregacja macierzy sztywności. Ponieważ macierze sztywności dla poszczególnych elementów zostały utworzone jako macierze 8×8, upraszcza się sposób uzyskiwania macierzy sztywności dla całego układu. Macierze sztywności w globalnym układzie odniesienia dla poszczególnych elementów należy zsumować.

$$[K^o] = [k^o_{1-2}] + [k^o_{1-3}] + [k^o_{2-3}] + [k^o_{2-4}] + [k^o_{3-4}]$$

W ten sposób można uzyskać globalną macierz sztywności jak poniżej

$$[K^o] = 10^6 \cdot \begin{bmatrix} 26{,}8 & 5{,}58 & -11{,}2 & -5{,}58 & -15{,}6 & 0 & 0 & 0 \\ 5{,}58 & 2{,}79 & -5{,}58 & -2{,}79 & 0 & 0 & 0 & 0 \\ -11{,}2 & -5{,}58 & 22{,}3 & 0 & 0 & 0 & -11{,}2 & 5{,}58 \\ -5{,}58 & -2{,}79 & 0 & 36{,}8 & 0 & -31{,}2 & 5{,}58 & -2{,}79 \\ -15{,}6 & 0 & 0 & 0 & 31{,}2 & 0 & -15{,}6 & 0 \\ 0 & 0 & 0 & -31{,}2 & 0 & 31{,}2 & 0 & 0 \\ 0 & 0 & -11{,}2 & 5{,}58 & -15{,}6 & 0 & 26{,}8 & -5{,}58 \\ 0 & 0 & 5{,}58 & -2{,}79 & 0 & 0 & -5{,}58 & 2{,}79 \end{bmatrix}$$

Jak widać, uzyskana macierz sztywności układu w globalnym układzie odniesienia jest symetryczna.

Obciążenia są podane w globalnym układzie odniesienia. Wystarczy wpisać je w odpowiednie miejsce macierzy obciążeń, która przyjmuje postać

$$\{P^o_i\} = \left\{ \begin{array}{cccccccc} 0 & 0 & 0 & -10^4 & 0 & 0 & 0 & 0 \end{array} \right\}^T$$

Nałożenie warunków brzegowych wynikających ze sposobu podparcia kratownicy. Nakładanie warunków brzegowych Dirichleta realizuje się przez wyzerowanie

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wiersza i kolumny macierzy sztywności odpowiadających danemu stopniowi swobody. Na głównej przekątnej, na przecięciu wspomnianego wiersza i kolumny, pozostawia się wartość 1, zeruje się wektor obciążeń dla odpowiednich stopni swobody. Dzieje się tak, ponieważ odebranie stopnia swobody oznacza zerowe przemieszczenia w danym stopniu swobody. Po modyfikacji macierz $[K^o]$ przyjmuje postać

$$[K^o] = 10^6 \cdot \begin{bmatrix} 10^{-6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10^{-6} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 22{,}3 & 0 & 0 & 0 & -11{,}2 & 0 \\ 0 & 0 & 0 & 36{,}8 & 0 & -31{,}2 & 5{,}58 & 0 \\ 0 & 0 & 0 & 0 & 31{,}2 & 0 & -15{,}6 & 0 \\ 0 & 0 & 0 & -31{,}2 & 0 & 31{,}2 & 0 & 0 \\ 0 & 0 & -11{,}2 & 5{,}58 & -15{,}6 & 0 & 26{,}8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10^{-6} \end{bmatrix}$$

Macierz obciążeń $\{P_i^o\}$ nie ulega zmianie, gdyż wiersze 1, 2 i 8 zawierają zera.

Rozwiązanie układu równań można zrealizować z wykorzystaniem MATLAB-a, GNU Octave lub innego środowisko obliczeniowego. Jako wynik otrzymuje się wektor przemieszczeń węzłowych

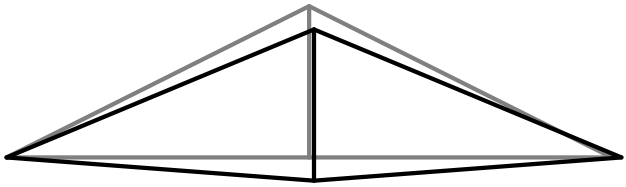
$$\{u_i^o\} = [K^o]^{-1}\{P_i^o\}$$

Przemieszczenia poszczególnych węzłów są przedstawione poniżej

$$\{u_i^o\} = 10^{-4} \begin{Bmatrix} 0{,}00 \\ 0{,}00 \\ 6{,}41 \\ -30{,}74 \\ 6{,}41 \\ -30{,}74 \\ 12{,}82 \\ 0{,}00 \end{Bmatrix} \begin{matrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ x_3 \\ y_3 \\ x_4 \\ y_4 \end{matrix}$$

Zazwyczaj łatwiej jest ocenić rozwiązanie, gdy dokona się wizualizacji (patrz rysunek 11.13, na którym przemieszczenia węzłów powiększono 50 razy), co pozwala na odrzucenie zdecydowanie błędnych rozwiązań.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.13. Wizualizacja odkształceń kratownicy

Po wyznaczeniu przemieszczeń węzłów można przystąpić do wyznaczania naprężeń w poszczególnych elementach w następujący sposób:

— wyznaczenie odkształceń

$$\varepsilon = [B]\{u_i^o\},$$

— wyznaczenie sił wewnętrznych

$$N_w = EA\varepsilon = EA[B]\{u_i^o\},$$

— wyznaczenie naprężeń

$$\sigma = \frac{N_w}{A} = E[B]\{u_i^o\}.$$

Niestety w przypadku elementów dwuwęzłowych z liniowymi funkcjami kształtu naprężenia w całym elemencie mają identyczną wartość. Aby badać rozkład naprężeń wzdłuż długości pręta, należy użyć elementu z innymi funkcjami kształtu.

11.10. Zagadnienia dynamiczne

Metoda elementów skończonych znajduje zastosowanie nie tylko do rozwiązywania problemów statyki. Metoda pozwala także na rozwiązywanie problemów dynamicznych. Równanie opisujące dynamikę układu z tłumieniem proporcjonalnym do prędkości, zdyskretyzowanego elementami skończonymi, przedstawia się następująco

$$[m]\{\ddot{u}_i\} + [c]\{\dot{u}_i\} + [k]\{u_i\} = \{P_i\} \tag{11.83}$$

gdzie kropkami nad zmienną oznaczono pochodną względem czasu, czyli

$$\{\ddot{u}_i\} = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\{u_i\}$$

$$\{\dot{u}_i\} = \frac{\mathrm{d}}{\mathrm{d}t}\{u_i\}$$

Jak pokazano wcześniej, macierze bezwładności i sztywności dla założonego rozkładu przemieszczeń można wyznaczyć bez większych problemów, natomiast najwięcej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

problemów sprawia macierz tłumienia. Postać tej macierzy w przypadku tłumienia materiałowego można wyznaczyć na podstawie zależności

$$[c] = \int_{\Omega} [N]^T \mu [N] \mathrm{d}\Omega$$

Oprócz tłumienia materiałowego spotyka się tłumienie wiskotyczne czy suche. W sposób przybliżony można opisywać tłumienie, charakteryzując ilość energii rozpraszanej w trakcie okresu drgań, i wyznaczać zastępczy współczynnik tłumienia wiskotycznego.

Z uwagi na złożoność zjawiska rozpraszania energii i trudności w dokładnym jego opisie, często przyjmuje się tzw. proporcjonalne tłumienie (zwane też tłumieniem Rayleigha), czyli macierz tłumienia wyraża się za pomocą macierzy sztywności i bezwładności

$$[c] = \epsilon_0 [m] + \epsilon_1 [k],$$

gdzie współczynniki ϵ_0 i ϵ_1 są współczynnikami Rayleigha. Taki sposób modelowania tłumienia ma zasadnicze zalety, ponieważ współczynniki Rayleigha można wyznaczać dla każdej postaci drgań własnych oddzielnie, zaś same macierze tłumienia stają się diagonalne. Spośród całej gamy zagadnień dynamicznych rozwiązywanych za pomocą metody elementów skończonych, jednym z częściej spotykanych jest zagadnienie własne, czyli wyznaczenie częstości drgań własnych układu. Zazwyczaj wartości z macierzy tłumienia są najtrudniejsze do wyznaczenia. Z tego powodu, a także z powodu trudności rozwiązania problemu z tłumieniem oraz niewielkimi różnicami częstości własnych nietłumionych i częstości własnych tłumionych, tłumienia się nie uwzględnia. Problem własny dla układu bez tłumienia można postawić w następujący sposób

$$[m]\{\ddot{u}_i\} + [k]\{u_i\} = 0,$$

co dla rozwiązania harmonicznego $u_i = Ae^{i\omega t}$ daje

$$\{\ddot{u}_i\} = -\omega^2 \{u_i\},$$

czyli równanie ruchu przyjmuje postać

$$\left([k] - \omega^2 [m]\right) \{u_i\} = 0. \tag{11.84}$$

Rozwiązanie tego problemu można sprowadzić do znalezienia pierwiastków równania charakterystycznego

$$\left|[k] - \omega^2 [m]\right| = 0$$

i wyznaczenia postaci drgań własnych $\{u_i\}$ oraz częstości drgań własnych ω. Po pomnożeniu równania (11.84) lewostronnie przez $[m]^{-1}$ można zapisać

$$\left([m]^{-1}[k] - \omega^2 \mathbb{I}\right) \{u_i\} = 0 \tag{11.85}$$

Można zastosować podstawienia $\lambda = \omega^2$, $\mathbb{A} = [m]^{-1}[k]$, $\mathbb{X} = \{u_i\}$, uzyskując następującą, dobrze znaną z matematyki, postać problemu znajdowania wartości i wektorów własnych

$$(\mathbb{A} - \lambda \mathbb{I})\mathbb{X} = 0$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po rozwiązaniu tak sformułowanego zagadnienia wartości własne λ_i stanowią kwadraty częstości własnych. Wektory własne $\mathbb{X}_i$ przedstawiają postacie drgań odpowiadające odpowiednim wartościom własnym.

Rozwiązanie przedstawionego problemu można osiągnąć różnymi sposobami. Poniżej przedstawiony zostanie jeden z nich. Po przyjęciu $\lambda = \omega^2$, równanie (11.85) można zapisać w następującej postaci

$$[m]^{-1}[k]\{u_i\} = \lambda\{u_i\}$$

Jeżeli macierz $[m]$ jest rzeczywista, symetryczna i dodatnio określona, to można wykorzystać rozkład Cholesky'ego i zapisać $[m] = LL^T$, gdzie L jest macierzą trójkątną dolną. W tej sytuacji można zapisać $[m]^{-1} = (L^T)^{-1}L^{-1}$, co daje

$$(L^T)^{-1}L^{-1}[k]\{u_i\} = \lambda\{u_i\}$$

Po pomnożeniu lewostronnym przez L^T otrzyma się

$$L^{-1}[k]\{u_i\} = \lambda L^T\{u_i\}$$

Niech teraz $x = L^T\{u_i\}$, czyli $\{u_i\} = (L^T)^{-1}x$, co daje

$$L^{-1}[k](L^T)^{-1}x = \lambda x$$

czyli można napisać następującą zależność

$$\lambda = \omega^2 = L^{-1}[k](L^T)^{-1} \tag{11.86}$$

Podobne rozumowanie można przeprowadzić, gdy to macierz sztywności rozłoży się na macierze trójkątne, przy czym w takim przypadku $\lambda = \frac{1}{\omega^2}$. Wartości własne wyznaczają częstości drgań własnych, ale dla wektorów własnych jest inaczej. Ponieważ wektory własne, czyli postacie drgań własnych wyznacza się z dokładnością do mnożnika, nie są istotne wartości przypisane poszczególnym węzłom. Ważne są stosunki tych wartości. Zatem można, a nawet należy dokonać normalizacji tych wektorów. Zazwyczaj normalizuje się je do jedynki lub względem masy. Przy normalizacji do jedynki postępuje się według następującej zależności

$$\{u_i\}_N = \frac{\{u_i\}}{\max(u_i)} \tag{11.87}$$

W praktyce często stosuje się normalizację względem masy. Przy normalizacji względem masy wykorzystywane jest równanie

$$\{u_i\}_j^T[m]\{u_i\}_j = 1 \tag{11.88}$$

gdzie $\{u_i\}_j$ jest j-tym wektorem własnym.

Normalizacja wektorów własnych względem masy ma pewne korzystne właściwości, które są wykorzystywane przy rozwiązywaniu zagadnień dynamicznych metodą superpozycji modalnej.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

11.10.1. Superpozycja modalna

Superpozycja modalna służy do rozwiązywania równań dynamiki układu dyskretnego (11.83), które dla ciągłości wywodu zostanie powtórzone poniżej

$$[m]\{\ddot{u}_i\} + [c]\{\dot{u}_i\} + [k]\{u_i\} = \{P_i\}$$

Aby można było rozwiązać takie równanie, konieczne jest zadanie warunków początkowych $\{u_i\}_0 = \{u_0\}$ i $\{\dot{u}_i\}_0 = \{\dot{u}_0\}$.

Równanie to w istocie stanowi układ sprzężonych równań różniczkowych zwyczajnych rzędu drugiego. Można próbować rozwiązać to równanie, „składając" rozwiązanie z funkcji ortogonalnych uzyskanych z rozwiązania problemu własnego w postaci

$$[m]\{\ddot{u}_i\} + [k]\{u_i\} = 0 \tag{11.89}$$

Po rozwiązaniu problemu własnego uzyskuje się zbiór N postaci drgań własnych, zwanych też wektorami własnymi, które są kolumnowymi wektorami $\{u_i\}_j$. Można je ułożyć w macierzy

$$\mathbf{\Phi} = \begin{bmatrix} \{u_i\}_1 & \{u_1\}_2 & \dots & \{u_i\}_N \end{bmatrix} \tag{11.90}$$

Ważne jest, by wektory własne były skalowane do jedynki względem masy, co nazywane jest M-ortonormalnością, czyli

$$\{u_i\}_j^T [m] \{u_i\}_j = 1$$

Podobnie można ułożyć w diagonalnej macierzy $\mathbf{\Omega}$ wartości własne $\lambda = \omega^2$.

$$\mathbf{\Omega}^2 = \begin{bmatrix} \omega_1^2 & 0 & \dots & 0 \\ 0 & \omega_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_N^2 \end{bmatrix} \tag{11.91}$$

Częstości drgań własnych, a co za tym idzie, postacie drgań, powinny być uporządkowane w kolejności rosnącej $\omega_1 < \omega_2 < \ldots < \omega_N$. Zakładając następującą postać rozwiązania harmonicznego $\{u_i\} = \{A\}e^{j\omega t}$, można wyrazić równanie (11.89) jako

$$-\omega^2 [m]\{u_i\} + [k]\{u_i\} = 0 \tag{11.92}$$

Równanie (11.92) można zapisać dla pojedynczego wektora własnego i odpowiadającej mu wartości własnej, czyli postaci i częstości drgań własnych, jako

$$[k]\{u_i\}_j - \omega^2 [m]\{u_i\}_j = 0 \tag{11.93}$$

zaś dla wszystkich wektorów własnych i wartości własnych, za pomocą zdefiniowanych wcześniej macierzy, w postaci

$$[k]\mathbf{\Phi} = [m]\mathbf{\Phi}\mathbf{\Omega}^2 \tag{11.94}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Z kolei równanie M-ortonormalności można zapisać w postaci

$$\mathbf{\Phi}^T[m]\mathbf{\Phi} = \mathbb{I} \tag{11.95}$$

gdzie $\mathbb{I}$ oznacza macierz jednostkową. Pomnożenie prawostronne obu stron powyższej zależności przez $\mathbf{\Omega}^2$ daje

$$\mathbf{\Phi}^T \underbrace{[m]\mathbf{\Phi}\mathbf{\Omega}^2}_{[k]\mathbf{\Phi}} = \mathbf{\Omega}^2 \tag{11.96}$$

czyli

$$\mathbf{\Phi}^T[k]\mathbf{\Phi} = \mathbf{\Omega}^2 \tag{11.97}$$

Powyższe równanie oznacza, że macierz sztywności również spełnia warunek ortogonalności wektorów własnych. Zatem macierz $\mathbf{\Phi}$ można wykorzystać do transformacji układu równań różniczkowych opisujących dynamikę badanego układu. Po tej transformacji przemieszczenia węzłów będą się wyrażały następującą zależnością

$$\{u_i\}(\tau) = \mathbf{\Phi}\{x_i\}(\tau) \tag{11.98}$$

Oznacza ona, że składowe wektora $\{x_i\}$ są udziałami przemieszczeń od poszczególnych postaci drgań w przemieszczeniu całkowitym. Dzięki temu równania dynamiki układu (11.83) można zapisać w postaci

$$[m]\mathbf{\Phi}\{\ddot{x}_i\}(\tau) + [c]\mathbf{\Phi}\{\dot{x}_i\}(\tau) + [k]\mathbf{\Phi}\{x_i\}(\tau) = \{P_i\}(\tau) \tag{11.99}$$

Po lewostronnym pomnożeniu obu stron równania przez $\mathbf{\Phi}^T$ oraz, dla uproszczenia zapisu, pominięciu jawnej zależności od czasu, równanie przyjmie postać

$$\underbrace{\mathbf{\Phi}^T[m]\mathbf{\Phi}}_{\mathbb{I}}\{\ddot{x}_i\} + \mathbf{\Phi}^T[c]\mathbf{\Phi}\{\dot{x}_i\} + \underbrace{\mathbf{\Phi}^T[k]\mathbf{\Phi}}_{\mathbf{\Omega}^2}\{x_i\} = \mathbf{\Phi}^T\{P_i\} \tag{11.100}$$

Po zastosowaniu odpowiednich, wcześniej wprowadzonych związków, rozwiązywane równanie można zapisać jako

$$\{\ddot{x}_i\} + \mathbf{\Phi}^T[c]\mathbf{\Phi}\{\dot{x}_i\} + \mathbf{\Omega}^2\{x_i\} = \mathbf{\Phi}^T\{P_i\} \tag{11.101}$$

Zgodnie z zależnością (11.98) przekształca się warunki początkowe. Przyjmują one postać

$$\{u_i\}(\tau_0) = \{u_0\}, \quad \text{czyli } \mathbf{\Phi}\{x_i\}(\tau_0) = \{u_0\} \tag{11.102}$$

Mnożąc lewostronnie przez $\mathbf{\Phi}^T[m]$, otrzymuje się zależność pozwalającą wyrazić warunek początkowy w przestrzeni przemieszczeń modalnych.

$$\underbrace{\mathbf{\Phi}^T[m]\mathbf{\Phi}}_{\mathbb{I}}\{x_i\}(\tau) = \mathbf{\Phi}^T[m]\{u_0\} \tag{11.103}$$

W związku z tym warunki początkowe dotyczące przemieszczeń i prędkości przyjmą postać

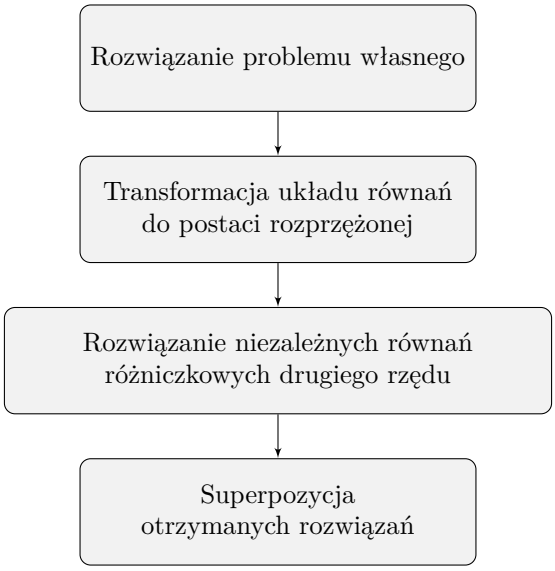
$$\{x_i\}(\tau) = \mathbf{\Phi}^T[m]\{u_0\}, \quad \text{oraz} \quad \{\dot{x}_i\}(\tau) = \mathbf{\Phi}^T[m]\{\dot{u}_0\} \tag{11.104}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Gdy w układzie nie uwzględnia się tłumienia, równanie (11.101) można zapisać w postaci układu rozprzężonych równań różniczkowych zwyczajnych rzędu drugiego

$$\ddot{x}_j + \omega_j^2 x_j = \{u_i\}_j^T \{P_i\} \tag{11.105}$$

z warunkami początkowymi (11.104). Zatem można stwierdzić, iż złożony problem rozwiązania równania (11.83) da się rozwiązać w przestrzeni przemieszczeń modalnych znacznie prościej. Algorytm rozwiązywania problemu dynamiki układu za pomocą superpozycji modalnej przedstawiono na rysunku 11.14.


Rozwiązanie problemu własnego
Transformacja układu równań do postaci rozprzężonej
Rozwiązanie niezależnych równań różniczkowych drugiego rzędu
Superpozycja otrzymanych rozwiązań

Rysunek 11.14. Algorytm postępowania przy superpozycji modalnej

Pomimo pozornego skomplikowania postępowania, uzyskuje się możliwość znacznego ograniczenia nakładów obliczeniowych koniecznych do rozwiązania problemu. Zasadniczą zaletą superpozycji modalnej jest możliwość rozwiązania tylko pewnego niewielkiego podzbioru równań opisujących zachowanie się obiektu. Spośród N równań wybiera się K związanych z najniższymi częstościami drgań własnych, przy czym najczęściej zachodzi zależność $K \ll N$.

Aby uwzględnić tłumienie w modelu modalnym, przyjmuje się, że macierz tłumienia również spełnia warunek M-ortogonalności, czyli

$$\{u_i\}_j^T [c] \{u_i\}_k = 0, \quad \text{dla} \quad j \neq k$$

Dla szczególnego przypadku, gdy $j = k$, niech zachodzi równość

$$\{u_i\}_j^T [c] \{u_i\}_j = 2\omega_j \xi_j$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie ξ_j, dla $j = 1, 2, \ldots, N$ są parametrami tłumienia modalnego. Zatem rozprzężone równania ruchu można przedstawić w postaci

$$\ddot{x}_j + 2\omega_j \xi_j \dot{x}_j + \omega_j^2 x_j = \{u_i\}_j^T \{P_i\} \tag{11.106}$$

Warunek M-ortogonalności dla macierzy tłumienia spełniony może być na przykład wtedy, gdy macierz tłumienia jest kombinacją liniową macierzy bezwładności i sprężystości.

$$[c] = \epsilon_0 [m] + \epsilon_1 [k]$$

Postaci macierzy tłumienia spełniającej warunek M-ortogonalności jest nieskończenie wiele [26]. Taka postać macierzy tłumienia pozwala uwzględnić różne rodzaje tłumienia występujące w modelowanych obiektach, wymaga jedynie identyfikacji współczynników ϵ_0 oraz ϵ_1. Zazwyczaj konieczne jest wykonanie eksperymentu modalnego, by prawidłowo określić wartości tych współczynników. Jak już wcześniej wspominano, jest to tak zwane tłumienie Rayleigha. Aby rozwiązać uzyskany rozprzężony układ równań, można wykorzystać którąś z opisanych metod rozwiązywania równań różniczkowych zwyczajnych rzędu drugiego lub wykorzystać całkę Duhamela. Gdy w równaniu nie uwzględnia się tłumienia, tak jak w równaniu (11.105), całka Duhamela przyjmuje postać

$$x_j(\tau) = \frac{1}{\omega_j} \int_{t_0}^{\tau} \{u_i\}_j^T \{P_i\} \sin\left(\omega_j(\tau - \kappa)\right) \mathrm{d}\kappa + A_j \sin \omega_j \tau + B_j \cos \omega_j \tau \tag{11.107}$$

Stałe A oraz B wyznacza się na podstawie warunków początkowych.

Gdy w równaniu uwzględnia się tłumienie (równanie (11.106)), całka Duhamela przyjmuje postać

$$x_j(\tau) = \frac{1}{\omega_j} \int_{t_0}^{\tau} \{u_i\}_j^T \{P_i\} e^{-\xi_j \omega_j (\tau - \kappa)} \sin\left(\bar{\omega}_j(\tau - \kappa)\right) \mathrm{d}\kappa + \\ + e^{-\xi_j \omega_j \tau} \left(A_j \sin \bar{\omega}_j \tau + B_j \cos \bar{\omega_j} \tau\right) \tag{11.108}$$

gdzie $\bar{\omega}_j = \omega_j \sqrt{1 - \xi_j^2}$ jest częstością drgań własnych tłumionych.

Po rozwiązaniu rozprzężonych równań różniczkowych opisujących zachowanie układu dla poszczególnych postaci drgań i częstości drgań, można dokonać superpozycji rozwiązań i uzyskać przybliżone rozwiązanie.

11.10.2. Przykład analizy drgań swobodnych pręta

Sformułowanie problemu. Wyznaczyć wartość pierwszej częstości drgań własnych stalowego pręta utwierdzonego lewostronnie (patrz rys. 11.15), o długości $l = 1$ m. Przyjąć wartości parametrów materiałowych $\rho = 7{,}65$ kg/m^3, $E = 2{,}1 \cdot 10^5$ MPa oraz pole przekroju poprzecznego $A = 25$ cm^2. Ocenić dokładność rozwiązania w zależności od liczby elementów przyjętych do dyskretyzacji. Wykorzystać elementy liniowe, paraboliczne i sześcienne.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

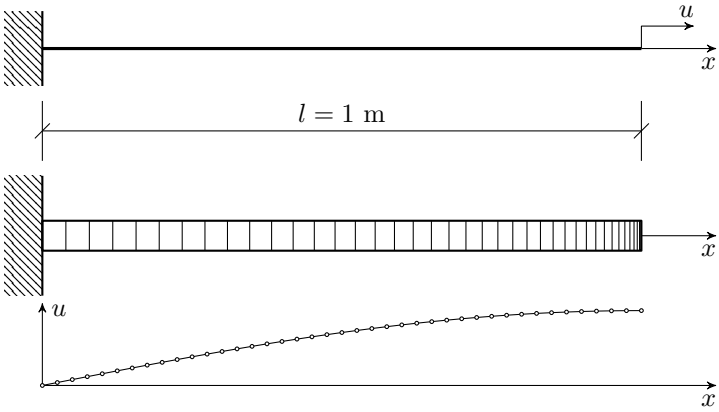
Rozwiązanie. Schemat analizowanego problemu przedstawia rysunek 11.15. W dolnej części rysunku przedstawione zostały schematycznie zaznaczone przemieszczenia przekrojów, które pojawiają się, gdy pręt drga z pierwszą częstością drgań własnych. Opisywane są one przez postać drgań własnych. Na podstawie zależności analitycznych, można wyznaczyć pierwszą częstość drgań własnych utwierdzonego pręta, która wynosi

$$\omega_1 = \frac{\pi}{2l}\sqrt{\frac{E}{\rho}} = \frac{\pi}{2 \cdot 1}\sqrt{\frac{2{,}1 \cdot 10^{11}}{7{,}65 \cdot 10^3}} = 8230\,\frac{\text{rad}}{\text{s}}$$

Tok obliczeń MES z elementami liniowymi. Aby wyznaczyć częstość drgań własnych pręta, należy:

— dokonać dyskretyzacji,
— wyznaczyć macierze sztywności dla elementów,
— wyznaczyć macierze bezwładności dla elementów,
— dokonać agregacji macierzy sztywności,
— dokonać agregacji macierzy bezwładności,
— nałożyć warunki brzegowe Dirichleta, co oznacza odebranie stopni swobody w węźle odpowiadającym za utwierdzenie,
— rozwiązać równanie (11.85).

Agregacja macierzy sztywności przebiega dokładnie jak w przykładzie z podrozdziału 11.7.2. Identycznie przebiega agregacja macierzy bezwładności.



Rysunek 11.15. Schemat drgań wzdłużnych pręta

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Dla modelu z trzema elementami skończonymi macierze przyjmują postać następującą

$$[k] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3{,}15 \cdot 10^9 & -1{,}575 \cdot 10^9 & 0 \\ 0 & -1{,}575 \cdot 10^9 & 3{,}15 \cdot 10^9 & -1{,}575 \cdot 10^9 \\ 0 & 0 & -1{,}575 \cdot 10^9 & 1{,}575 \cdot 10^9 \end{bmatrix}$$

$$[m] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4{,}25 & 1{,}0625 & 0 \\ 0 & 1{,}0625 & 4{,}25 & 1{,}0625 \\ 0 & 0 & 1{,}0625 & 2{,}125 \end{bmatrix}$$

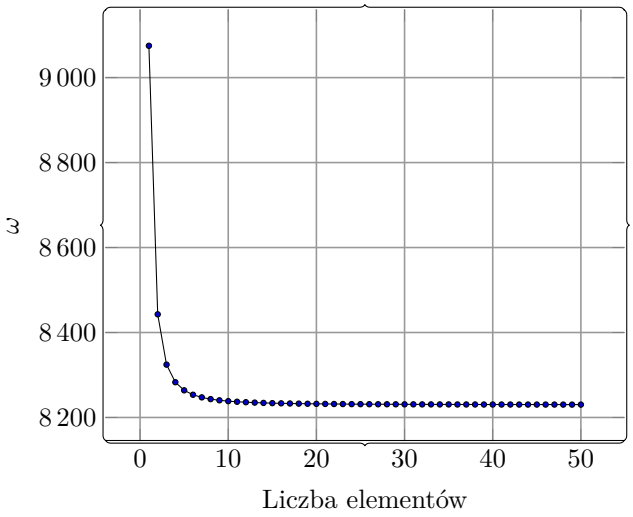
Wyniki, w zależności od liczby elementów skończonych liniowych, zostały przedstawione na rysunku 11.16. Widoczna jest stosunkowo niewielka zmiana poszukiwanej wartości pierwszej częstości drgań własnych od około 10 elementów. Można się przekonać, że dopiero przyjęcie 111 elementów skończonych liniowych daje wartość poszukiwanej częstości drgań własnych z dokładnością lepszą niż błąd maszynowy. Oczywiście należy sobie zdawać sprawę z tego, że w metodzie elementów skończonych błędy zależą nie tylko od dyskretyzacji. Nie zmienia to faktu, że powyżej 10 elementów zysk w postaci dokładności rozwiązania w porównaniu z kosztem obliczeniowym jest naprawdę niewielki. Z tego też powodu przy rozwiązywaniu problemów należy dokonać analizy zbieżności wyników w zależności od gęstości siatki. Dla kilku gęstości należy sprawdzić zmiany rozwiązania i ocenić, czy dalsze zagęszczanie siatki jest ekonomicznie uzasadnione.

Elementy paraboliczne. Tok obliczeń jest dokładnie taki sam jak dla elementów liniowych. Przy czym macierze sztywności i bezwładności mają postać jak w zależnościach (11.77) oraz (11.78). Zasadniczą korzyść z wykorzystania elementów parabolicznych widać z tabeli 11.2, przedstawiającej zależność między liczbą elementów skończonych a wartością pierwszej częstości drgań własnych wyznaczoną z modelu.

Jak widać z tabeli 11.2, przy modelowaniu pręta elementami trójwęzłowymi, czyli z paraboliczną funkcją kształtu, znacznie dokładniej jest wyznaczana wartość pierwszej częstości drgań własnych. Wynika to z faktu, że paraboliczny rozkład prędkości drgań wzdłuż długości pręta znacznie dokładniej oddaje rzeczywisty rozkład sinusoidalny niż funkcja liniowa.

Elementy sześcienne. Podobnie można wyznaczyć pierwszą częstość drgań własnych analizowanego pręta, wykorzystując elementy z funkcjami kształtu stopnia trzeciego. Macierze sztywności i bezwładności są opisane wzorami (11.80) oraz (11.81).

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 11.16. Wartość pierwszej częstości drgań własnych pręta w funkcji liczby elementów z liniowymi funkcjami kształtu

Tabela 11.2. Wartość pierwszej częstości drgań własnych w zależności od liczby elementów parabolicznych

Liczba elementów	ω [rad/s]	Błąd [%]
1	8260,9	0,375
2	8232,1	0,026
3	8230,4	0,005
4	8230,1	0,001
5	8230,0	–

Po zdyskretyzowaniu pręta za pomocą jednego elementu trzeciego stopnia, pierwsza częstość drgań własnych została obliczona jako $\omega_1 = 8230{,}5$ rad/s co daje błąd około 0,007%.

Tabela 11.3 przedstawia wyniki obliczeń dla pierwszej częstości drgań własnych pręta po dyskretyzacji elementami sześciennymi.

Wnioski Dla elementów liniowych podział na dwa elementy daje dokładność wyznaczenia pierwszej częstości drgań własnych około 3%. Podział na trzy elementy daje dokładność rzędu 1%.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 11.3. Wartość pierwszej częstości drgań własnych w zależności od liczby elementów sześciennych

Liczba elementów	ω [rad/s]	Błąd [%]
1	8230,5	0,007
2	8230,0	0,0002
3	8230,0	$9 \cdot 10^{-5}$

Jednocześnie widać, iż dla elementów liniowych dopiero podział na 111 elementów skończonych dał dokładną wartość pierwszej częstości drgań własnych, podczas gdy podział na 5 elementów parabolicznych daje dokładną wartość pierwszej częstości drgań własnych. Podstawowym wnioskiem, w zasadzie oczywistym, jest konieczność wykorzystywania elementów możliwie wysokiego stopnia. Daje to znaczne korzyści obliczeniowe. Opisane postępowanie dla jednego elementu skończonego jest analogiem metody Rayleigha wyznaczania częstości drgań własnych,która polega na wyznaczaniu energii kinetycznej i potencjalnej układu w skrajnym położeniu podczas drgań przy założonej postaci drgań.

11.10.3. Przykład analizy dynamicznej

Sformułowanie problemu. Dla kratownicy z przykładu zawartego w podrozdziale 11.9.1 wyznaczyć częstości drgań własnych oraz postacie drgań dla poszczególnych częstości. Wyliczone postacie drgań własnych przedstawić graficznie. Porównać wektory własne skalowane do jedności ze skalowanymi względem macierzy bezwładności.

Dane. Przyjąć pola przekrojów poszczególnych prętów oraz moduły Younga identyczne z wartościami przyjętymi w przykładzie 11.9.1. Oprócz wartości przyjętych w przywołanym przykładzie, należy określić gęstość materiału poszczególnych prętów. Przyjęto gęstość $\rho = 7\,860$ kg/m^3.

Tok obliczeń stanowi rozwinięcie analizy statycznej przeprowadzonej w poprzednim przykładzie.

W trakcie analizy statycznej wyznaczono macierze sztywności poszczególnych elementów oraz macierz sztywności całego układu. Określono też wygląd macierzy sztywności po nałożeniu warunków Dirichleta.

Dalsza część analizy sprowadza się do wyznaczenia macierzy bezwładności elementów według zależności (11.55), zagregowaniu ich do globalnej macierzy bezwładności oraz nałożeniu warunków brzegowych. Następnie należy wyznaczyć wartości i wek-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

tory własne zgodnie z zależnością (11.84). Zagregowana macierz bezwładności ma następującą postać

$$[m] = 10^{-2} \begin{bmatrix} 4{,}3284 & 0 & 1{,}1424 & 0 & 1{,}0218 & 0 & 0 & 0 \\ 0 & 4{,}3284 & 0 & 1{,}1424 & 0 & 1{,}0218 & 0 & 0 \\ 1{,}1424 & 0 & 5{,}5914 & 0 & 0{,}5109 & 0 & 1{,}1424 & 0 \\ 0 & 1{,}1424 & 0 & 5{,}5914 & 0 & 0{,}5109 & 0 & 1{,}1424 \\ 1{,}0218 & 0 & 0{,}5109 & 0 & 5{,}1090 & 0 & 1{,}0218 & 0 \\ 0 & 1{,}0218 & 0 & 0{,}5109 & 0 & 5{,}1090 & 0 & 1{,}0218 \\ 0 & 0 & 1{,}1424 & 0 & 1{,}0218 & 0 & 4{,}3284 & 0 \\ 0 & 0 & 0 & 1{,}1424 & 0 & 1{,}0218 & 0 & 4{,}3284 \end{bmatrix}$$

Warunki brzegowe zostaną uwzględnione podobnie jak w przypadku macierzy sztywności. Oczywiście, zamiast zerować odpowiednie wiersze i kolumny macierzy bezwładności, można je po prostu wykreślić, uzyskując korzyść w postaci mniejszej macierzy. Po uwzględnieniu warunków brzegowych macierz bezwładności wygląda następująco

$$[m] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}0559143 & 0 & 0{,}005109 & 0 & 0{,}0114241 & 0 \\ 0 & 0 & 0 & 0{,}0559143 & 0 & 0{,}005109 & 0 & 0 \\ 0 & 0 & 0{,}005109 & 0 & 0{,}05109 & 0 & 0{,}010218 & 0 \\ 0 & 0 & 0 & 0{,}005109 & 0 & 0{,}05109 & 0 & 0 \\ 0 & 0 & 0{,}0114241 & 0 & 0{,}010218 & 0 & 0{,}0432841 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Wykorzystując na przykład polecenia `eig(K,M)` czy `eig(inv(M)*K)`, stanowiące część pakietu MATLAB czy GNU Octave, można wyznaczyć wartości wektorów i wartości

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

własnych. Wyznaczone wektory własne zawiera macierz

$$\{u_i\} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0{,}241582 & -0{,}242397 & -0{,}173937 & 0{,}525037 & -0{,}734885 & 0 & 0 & 0 \\ -0{,}468496 & -0{,}436151 & 0{,}634139 & 0{,}266372 & 0{,}004291 & 0 & 0 & 0 \\ 0{,}400904 & -0{,}435435 & -0{,}168635 & 0{,}426531 & 0{,}64163 & 0 & 0 & 0 \\ 0{,}433454 & 0{,}484155 & 0{,}661724 & 0{,}348444 & 0{,}0365189 & 0 & 0 & 0 \\ -0{,}611182 & 0{,}571845 & -0{,}318273 & 0{,}591642 & 0{,}216574 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Wartości własne wyznaczone numerycznie przedstawiono w wektorze λ

$$\lambda = \begin{bmatrix} 1{,}42474 \cdot 10^8 \\ 1{,}27575 \cdot 10^8 \\ 2{,}32313 \cdot 10^6 \\ 1{,}33626 \cdot 10^7 \\ 5{,}32672 \cdot 10^7 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \qquad f = \begin{bmatrix} 1899{,}71 \\ 1797{,}64 \\ 242{,}581 \\ 581{,}789 \\ 1161{,}58 \\ 0{,}1592 \\ 0{,}1592 \\ 0{,}1592 \end{bmatrix} [Hz], \qquad f_{sort} = \begin{bmatrix} 242{,}581 \\ 581{,}789 \\ 1161{,}58 \\ 1797{,}64 \\ 1899{,}71 \end{bmatrix}$$

Najwygodniej jest posortować częstości drgań własnych rosnąco oraz wyeliminować te wartości, które wynikają ze sposobu uwzględniania warunków brzegowych, czyli odbierania stopni swobody. Te wartości własne, które przyjmują wartość równą jeden, wynikają z uwzględnienia warunków brzegowych, więc należy je pominąć. Posortowane częstotliwości drgań własnych zostały zebrane w macierzy f_{sort}, zaś odpowiadające im postacie drgań własnych przedstawiono na rysunku 11.17.

Ponieważ postacie drgań własnych opisywane są przez stosunki amplitud drgań, zatem wektory własne można przeskalować, przy zachowaniu proporcji między poszczególnymi składowymi wektora. Jak wspomniano wcześniej, można je skalować względem macierzy bezwładności lub do jedności. Po przeskalowaniu względem ma-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

cierzy bezwładności, zgodnie z zależnością (11.88) otrzymuje się następującą postać macierzy wektorów własnych.

$$\{u_i\} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1{,}2072 & -1{,}2059 & -0{,}7141 & 2{,}0512 & -3{,}3642 & 0 & 0 & 0 \\ -2{,}3412 & -2{,}1698 & 2{,}6033 & 1{,}0406 & 0{,}0196 & 0 & 0 & 0 \\ 2{,}0033 & -2{,}1663 & -0{,}6923 & 1{,}6663 & 2{,}9373 & 0 & 0 & 0 \\ 2{,}1660 & 2{,}4087 & 2{,}7166 & 1{,}3613 & 0{,}1672 & 0 & 0 & 0 \\ -3{,}0542 & 2{,}8449 & -1{,}3066 & 2{,}3114 & 0{,}9915 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

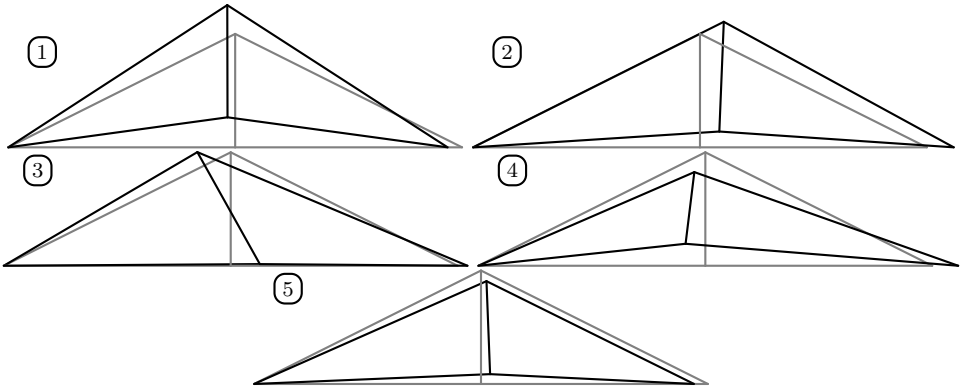
Wykorzystując zależność (11.87), można skalować składniki wektorów własnych tak, by maksymalna wartość wynosiła ± 1.

$$\{u_i\} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0{,}3952 & -0{,}4239 & -0{,}2629 & 0{,}8874 & -1 & 0 & 0 & 0 \\ -0{,}7665 & -0{,}7627 & 0{,}9583 & 0{,}4502 & 0{,}0058 & 0 & 0 & 0 \\ 0{,}6560 & -0{,}7615 & -0{,}2548 & 0{,}7209 & 0{,}8731 & 0 & 0 & 0 \\ 0{,}7092 & 0{,}8467 & 1 & 0{,}5889 & 0{,}0497 & 0 & 0 & 0 \\ -1 & 1 & -0{,}4810 & 1 & 0{,}2947 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Na podstawie wyliczonych wektorów własnych można narysować postacie drgań własnych. Postać drgań własnych charakteryzuje sposób poruszania się poszczególnych węzłów. Na rysunku 11.17 przedstawiono wygląd kolejnych postaci drgań własnych analizowanej kratownicy. Częstości i postacie drgań dają istotną informację o zachowaniu analizowanego obiektu pod wpływem wymuszeń, niezależnie od ich charakteru. Wykorzystując superpozycję modalną, można analizować zachowanie obiektu pod wpływem wymuszeń monoharmonicznych, poliharmonicznych czy wymuszeń impulsowych, a nawet szumowych.

Po omówieniu sposobu modelowania układów za pomocą elementów prętowych, zostanie przedstawiony sposób budowy macierzy bezwładności i sztywności elementów belkowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

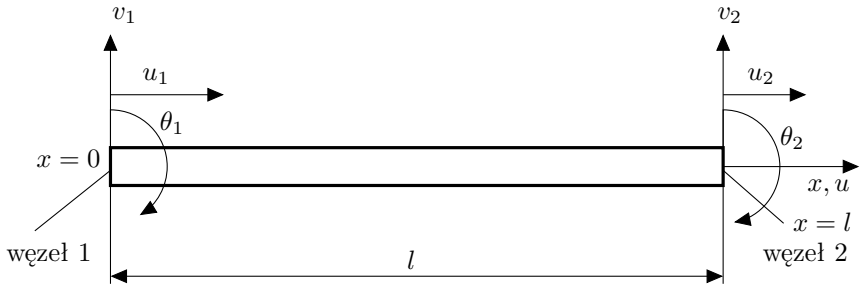


Rysunek 11.17. Postacie drgań własnych kratownicy wartości przemieszczeń zostały przeskalowane

Wykorzystany model to tak zwana belka Bernoulliego, dla której głównym założeniem jest zachowanie płaskości przekrojów belki przy odkształceniu, co jest równoznaczne z niewielkimi odkształceniami.

11.11. Macierz sztywności i bezwładności elementu belkowego

Element belkowy omawiany w tym podrozdziale różni się od prętowego tym, że ma cztery stopnie swobody oraz podlega wyłącznie zginaniu. Schemat elementu przedstawiono na rysunku 11.18.



Rysunek 11.18. Schemat zginanego elementu belkowego

W każdym węźle określa się przemieszczenie v oraz kąt obrotu θ. Między tymi stopniami swobody zachodzi związek (11.109).

$$\theta = -\frac{\mathrm{d}v}{\mathrm{d}x} \tag{11.109}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zmianę przemieszczeń i kątów obrotu wzdłuż elementu można interpolować na przykład wielomianem stopnia trzeciego

$$v = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \alpha_4 x^3 \tag{11.110}$$

zaś na podstawie (11.109) uzyska się

$$\theta = -\frac{dv}{dx} = -\alpha_2 - 2\alpha_3 x - 3\alpha_4 x^2 \tag{11.111}$$

Wyznaczenia współczynników α_i można dokonać na podstawie wartości brzegowych dla elementu w postaci układu równań (11.112).

$$\begin{cases} x = 0, & v = v_1, & \alpha_1 = v_1 \\ x = 0, & \theta = \theta_1, & \alpha_2 = -\theta_1 \\ x = l, & v = v_2, & v_2 = v_1 - \theta_1 l + \alpha_3 l^2 + \alpha_4 l^3 \\ x = l, & \theta = \theta_2, & \theta_2 = \theta_1 - 2\alpha_3 l - 3\alpha_4 l^2 \end{cases} \tag{11.112}$$

Po rozwiązaniu tego układu równań otrzymuje się wartości współczynników a_3 i a_4

$$\alpha_3 = 3\frac{v_2 - v_1}{l^2} + \frac{2\theta_1 + \theta_2}{l} \tag{11.113}$$

oraz

$$\alpha_4 = 2\frac{v_1 - v_2}{l^3} - \frac{\theta_1 + \theta_2}{l^2} \tag{11.114}$$

Po podstawieniu do równania (11.110) wielomian opisujący przemieszczenia wzdłuż belki wygląda następująco

$$v = v_1 - \theta_1 x + \left(3\frac{v_2 - v_1}{l^2} + \frac{2\theta_1 + \theta_2}{l}\right)x^2 + \left(2\frac{v_1 - v_2}{l^3} - \frac{\theta_1 + \theta_2}{l^2}\right)x^3$$

Tradycyjnie należy zastosować podstawienie $x = \xi l$. Po uporządkowaniu postać macierzowa równania jest następująca

$$v = \begin{bmatrix} 1 - 3\xi^2 + 2\xi^3 & l(-\xi + 2\xi^2 - \xi^3) & 3\xi^2 - 2\xi^3 & l(\xi^2 - \xi^3) \end{bmatrix} \begin{bmatrix} v_1 \\ \theta_1 \\ v_2 \\ \theta_2 \end{bmatrix} = [N]\{u_i\}$$

Widać, że wektor przemieszczeń $\{u_i\}$ wygląda następująco

$$\{u_i\} = \begin{bmatrix} v_1 & \theta_1 & v_2 & \theta_2 \end{bmatrix}^T$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

natomiast macierz funkcji kształtu ma postać

$$[N] = \begin{bmatrix} 1-3\xi^2+2\xi^3 & l(-\xi+2\xi^2-\xi^3) & 3\xi^2-2\xi^3 & l(\xi^2-\xi^3) \end{bmatrix}$$

Można sprawdzić, że funkcje kształtu stanowią wielomiany Hermite'a pierwszego rzędu. Aby wyznaczyć macierz sztywności przy zginaniu $[k]$ dla elementu belkowego należy rozważyć energię sprężystą zginania belki U_b

$$U_b = \int_{\Omega} \frac{M^2}{2EI} \mathrm{d}x$$

gdzie M jest momentem gnącym, E to moduł Younga, I jest momentem bezwładności przekroju belki względem osi obojętnej. Z porównania z równaniem (11.30) wynika, iż $\sigma = M$, $D = EI$, $\mathrm{d}\Omega = \mathrm{d}x$. Co więcej, należy pamiętać, że odkształceniem związanym z momentem gnącym jest krzywizna linii ugięcia, co stanowi drugą pochodną ugięcia względem zmiennej x

$$EI\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} = M$$

czyli

$$\{\varepsilon\} = \frac{\mathrm{d}^2 v}{\mathrm{d}x^2}$$

Zapisując energię zginania belki w postaci macierzowej, otrzymuje się

$$U_b = \frac{1}{2}\int_{\Omega} \{M\}^T \frac{1}{EI}\{M\}\mathrm{d}x = \frac{1}{2}\int_{\Omega} \left\{\frac{\mathrm{d}^2 v}{\mathrm{d}x^2}\right\} EI \left\{\frac{\mathrm{d}^2 v}{\mathrm{d}x^2}\right\} \mathrm{d}x$$

Pochodne przemieszczeń można obliczyć z następujących zależności

$$\frac{dv}{dx} = \frac{dv}{ld\xi} = \begin{bmatrix} \frac{1}{l}(-6\xi+6\xi^2) & -1+4\xi-3\xi^2 & \frac{1}{l}(6\xi-6\xi^2) & 2\xi-3\xi^2 \end{bmatrix} \begin{bmatrix} v_1 \\ \theta_1 \\ v_2 \\ \theta_2 \end{bmatrix}$$

oraz

$$\frac{d^2 v}{dx^2} = \frac{d^2 v}{l^2 d\xi^2} = \begin{bmatrix} \frac{1}{l^2}(-6+12\xi) & \frac{1}{l}(4-6\xi) & \frac{1}{l^2}(6-12\xi) & \frac{1}{l}(2-6\xi) \end{bmatrix} \{u_i\} = [B]\{u_i\}.$$

Widać, że macierz $[B]$ w tym przypadku oznacza macierz drugich pochodnych macierzy funkcji kształtu. A zatem macierz sztywności elementu można obliczyć jak poprzednio z zależności

$$[k] = \int_{\Omega} [B]^T [D][B]\mathrm{d}\Omega \tag{11.115}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po podstawieniu uzyskanych macierzy $[B]$ oraz $[D]$, wymnożeniu i scałkowaniu poszczególnych elementów macierzy otrzymuje się macierz sztywności elementu przy zginaniu wyrażoną w lokalnym układzie współrzędnych

$$[k] = EI \begin{array}{c} \begin{matrix} v_1 & \theta_1 & v_2 & \theta_2 \end{matrix} \\ \begin{bmatrix} \frac{12}{l^3} & \frac{-6}{l^2} & \frac{-12}{l^3} & \frac{-6}{l^2} \\ \frac{-6}{l^2} & \frac{4}{l} & \frac{6}{l^2} & \frac{2}{l} \\ \frac{-12}{l^3} & \frac{6}{l^2} & \frac{12}{l^3} & \frac{6}{l^2} \\ \frac{-6}{l^2} & \frac{2}{l} & \frac{6}{l^2} & \frac{4}{l} \end{bmatrix} \end{array} \begin{matrix} v_1 \\ \theta_1 \\ v_2 \\ \theta_2 \end{matrix}$$

Podobnie, po podstawieniu macierzy funkcji kształtu dla zginania do wzoru (11.55) można uzyskać macierz bezwładności dla zginania wyrażoną w lokalnym układzie współrzędnych

$$[m] = \frac{\rho A l}{420} \begin{array}{c} \begin{matrix} v_1 & \theta_1 & v_2 & \theta_2 \end{matrix} \\ \begin{bmatrix} 156 & -22l & 54 & 13l \\ -22l & 4l^2 & -13l & -3l^2 \\ 54 & -13l & 156 & 22l \\ 13l & -3l^2 & 22l & 4l^2 \end{bmatrix} \end{array} \begin{matrix} v_1 \\ \theta_1 \\ v_2 \\ \theta_2 \end{matrix}$$

We współrzędnych globalnych macierz sztywności elementu belkowego obróconego o kąt α można wyznaczyć ze znanej zależności

$$[k_b^o] = [DC]^T [k] [DC]$$

gdzie

$$[DC] = \begin{array}{c} \begin{matrix} u_1 & v_1 & \theta_1 & u_2 & v_2 & \theta_2 \end{matrix} \\ \begin{bmatrix} \cos\alpha & \sin\alpha & 0 & 0 & 0 & 0 \\ -\sin\alpha & \cos\alpha & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos\alpha & \sin\alpha & 0 \\ 0 & 0 & 0 & -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ \theta_1 \\ u_2 \\ v_2 \\ \theta_2 \end{matrix}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

jest macierzą kosinusów kierunkowych. Po podstawieniu, macierz sztywności elementu belkowego we współrzędnych globalnych przyjmuje następującą postać

$$[k_b^o] = EI \begin{array}{c} \begin{matrix} u_1 & v_1 & \theta_1 & u_2 & v_2 & \theta_2 \end{matrix} \\ \begin{bmatrix} \frac{12}{l^3}s^2 & & & & & \\ -\frac{12}{l^3}cs & \frac{12}{l^3}c^2 & & & & \\ \frac{6}{l^2}s & -\frac{6}{l^2}c & \frac{4}{l} & & & \\ -\frac{12}{l^3}s^2 & \frac{12}{l^3}cs & -\frac{6}{l^2}s & \frac{12}{l^3}s^2 & & \\ \frac{12}{l^3}cs & -\frac{12}{l^3}c^2 & \frac{6}{l^2}c & -\frac{12}{l^3}cs & \frac{12}{l^3}c^2 & \\ \frac{6}{l^2}s & -\frac{6}{l}c & \frac{2}{l} & -\frac{6}{l^2}s & \frac{6}{l^2}c & \frac{4}{l} \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ \theta_1 \\ u_2 \\ v_2 \\ \theta_2 \end{matrix}$$

Dla zwięzłości zapisu zastosowano oznaczenia $c = \cos\alpha$ oraz $s = \sin\alpha$ i pominięto symetryczną część macierzy. Podobnie macierz bezwładności we współrzędnych globalnych wyraża się zależnością

$$[m_b^o] = [DC]^T[m][DC]$$

czyli po podstawieniu i wymnożeniu można zapisać

$$[m_b^o] = \begin{array}{c} \begin{matrix} u_1 & v_1 & \theta_1 & u_2 & v_2 & \theta_2 \end{matrix} \\ \begin{bmatrix} 156s^2 & & & & & \\ -156cs & 156c^2 & & & & \\ 22ls & -22lc & 4l^2 & & & \\ 54s^2 & -54cs & 13ls & 156s^2 & & \\ -54cs & 54c^2 & -13lc & -156cs & 156c^2 & \\ -13ls & 13lc & -3l^2 & -22ls & 22lc & 4l^2 \end{bmatrix} \end{array} \begin{matrix} u_1 \\ v_1 \\ \theta_1 \\ u_2 \\ v_2 \\ \theta_2 \end{matrix}$$

Zastosowano podobne oznaczenia jak powyżej dla macierzy sztywności oraz pominięto część symetryczną macierzy bezwładności.

Ponieważ uzyskane zależności dotyczyły jedynie obciążenia belki siłami poprzecznymi, wywołującymi zginanie, konieczne jest uwzględnienie sił osiowych. Zakładając niezależność tych dwóch stanów odkształceń, można wykorzystać wyniki uzyskane dla pręta podlegającego jedynie rozciąganiu. Zatem poszukiwana macierz sprężystości będzie miała następującą postać

$$[k^o] = [k_b^o] + [k_r^o] \text{ oraz } [m^o] = [m_b^o] + [m_r^o]$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie indeks r oznacza element rozciągany, indeks b oznacza belkę zginaną.

$$[k_r^o] = \frac{AE}{l}\begin{bmatrix} c^2 & cs & 0 & -c^2 & -cs & 0 \\ cs & s^2 & 0 & -cs & -s^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -c^2 & -cs & 0 & c^2 & cs & 0 \\ -cs & -s^2 & 0 & cs & s^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[m_r^o] = \frac{\rho Al}{6}\begin{bmatrix} 2c^2 & & & & & \\ 2cs & 2s^2 & & & & \\ 0 & 0 & 0 & & & \\ c^2 & cs & 0 & 2c^2 & & \\ cs & s^2 & 0 & 2cs & 2s^2 & \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Elementy belkowe wyższych rzędów buduje się podobnie. Jako wzór może posłużyć postępowanie przy elemencie prętowym.

11.12. Podsumowanie i kilka słów o błędach

Metoda elementów skończonych, jako przybliżona metoda rozwiązywania równań różniczkowych cząstkowych, jest obarczona błędami. Inżynier powinien mieć świadomość przybliżonego charakteru rozwiązań problemów uzyskanych z tej metody. Podstawowa grupa błędów wynika ze wszystkich założeń idealizacyjnych prowadzących do uzyskania zależności analitycznych opisujących badane zjawiska. Te zależności, przedstawiane zazwyczaj jako równania różniczkowe cząstkowe lub równania całkowe, podlegają przybliżonemu rozwiązywaniu za pomocą MES. Zatem błędy wpływające na jakość wyniku można podzielić na następujące grupy:

— błędy idealizacji modelu – inżynier nie ma wpływu na założenia równań opisujących problem;
— błędy danych wejściowych – wartości przyjmowane jako warunki brzegowe, wartości współczynników materiałowych, założenia o izotropii materiału, itp.;
— błędy dyskretyzacji – skończone pola przemieszczeń uogólnionych eliminują znaczną część możliwych uogólnionych deformacji, brak zagęszczeń siatki w obszarach z dużymi gradientami analizowanych wartości;

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— błędy skończonej reprezentacji liczbowej w maszynach cyfrowych (zaokrągleń) – obliczenia nie są prowadzone w dziedzinie liczb rzeczywistych (czy zespolonych), a w ograniczonym ich podzbiorze;
— błąd wynikający z samej metody i zastosowanych elementów – ewentualny brak ciągłości pochodnych między elementami, itp.

Zasadniczą trudnością w używaniu metod przybliżonych jest niemożność określenia a priori dokładnego rozwiązania. Nie znając rozwiązania, nie można określić błędu. Istniejące oszacowania błędu pozwalają na orientacyjne określenie poprawności rozwiązania, nie dają zaś pewności, że rozwiązanie jest właściwe. Często można błędne rozwiązanie problemu wychwycić „na oko" – od razu widać, że wyniki nie spełniają podstawowych założeń co do przemieszczeń czy naprężeń. Niemniej można trafić na problem błędnie rozwiązany mimo pozornie „rozsądnych" wyników. Niebezpieczeństwo zajścia takiego zjawiska znacznie rośnie, gdy stosuje się elementy niedostosowane.

Należy pamiętać, iż metoda elementów skończonych daje jedynie przybliżone rozwiązania, i to tylko w węzłach siatki. Jakość rozwiązania zależy od jakości siatki i jej gęstości, przy czym pojęcie jakości siatki ujmuje między innymi proporcje długości boków elementów, różnice w wielkości sąsiednich elementów czy wielkość kątów między bokami elementu.

Wariacyjne sformułowanie powoduje, iż globalnie układ osiąga minimum np. energii potencjalnej, natomiast lokalnie nie muszą być spełnione warunki równowagi. Wielkości wyznaczane bezpośrednio są dokładniejsze niż wyznaczane pośrednio, np. przemieszczenia są dokładniejsze od odkształceń, a tym samym naprężeń.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 12

Obliczenia symboliczne

12.1. Wstęp

W latach siedemdziesiątych ubiegłego wieku w ramach badań nad sztuczną inteligencją rozpoczęto rozwijanie systemów algebry komputerowej (*Computer Algebra System*), w skrócie CAS. Pierwszymi z nich były Reeduce i Macsyma. Obecnie największe możliwości mają Mathematica i Maple, powszechnie używane przez naukowców i inżynierów. Systemy algebry komputerowej wspomagają wykonywanie obliczeń symbolicznych w matematyce i innych dyscyplinach naukowych. Zazwyczaj operują na funkcjach jednej i wielu zmiennych. Mają zaimplementowane algorytmy obliczeń macierzowych. Są w stanie rozwiązywać równania i układy równań liniowych, nieliniowych, różniczkowych i całkowych. Zazwyczaj mają również możliwość rysowania wykresów oraz eksportu wyrażeń do systemów składu tekstu np. TEX czy LATEX. Często mogą wykonywać obliczenia numeryczne, a wyniki podawać w zasadzie z arbitralną precyzją.

W rozdziale zostaną przybliżone dwa takie pakiety: Maxima i SageMath. Ich cechą wspólną jest dostępność źródeł, co pozwala na ewentualną modyfikację kodu i wprowadzanie nowych algorytmów. Nie bez znaczenia jest brak opłat licencyjnych, co dla studenta stanowi zasadniczą korzyść. Jednocześnie opanowanie ich obsługi ułatwi ewentualną przyszłą pracę w innych pakietach z rodziny CAS.

Należy mieć świadomość, że przedstawiony materiał stanowi niezwykle wąski wycinek możliwości omawianych pakietów. Zakres materiału wydaje się wystarczający, by czytelnik opanował umiejętności przydatne w innych częściach niniejszego skryptu.

12.2. Maxima

Maxima[1] jest otwartym systemem CAS, ma różne odmiany różniące się interfejsem, takie jak wxMaxima, Xmaxima czy Emacs. Kwestią osobistą jest ocena wygody korzystania z różnych interfejsów. Składnia poleceń i podane przykłady działają niezależnie od używanego interfejsu.

Pracę z Maximą rozpoczyna się, wydając polecenie `maxima`, `xmaxima`, `texmacs` lub inne zależnie od systemu operacyjnego i wybranego interfejsu.

[1] Przykłady przetestowano w pakiecie Maxima w wersji 5.37.2

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Szczególnie przydatną informacją przy poznawaniu nowego oprogramowania jest sposób uzyskiwania pomocy na temat poleceń. Najłatwiej uzyskać ją, używając polecenia składającego się ze znaku zapytania oraz tekstu opisującego zagadnienie na przykład `? integrate` czy `? sum`. Zdecydowanie szerszy zakres wyszukiwania pomocy zapewnia operator `??` na przykład `?? diff`.

Zmienna specjalna % zawiera wynik ostatniego polecenia i może służyć do dalszych działań. Oprócz niej, Maxima tworzy zmienne %in oraz %on, gdzie n stanowi liczbę naturalną. Zmiennym %in są przypisywane odpowiednio kolejne wprowadzone wyrażenia, zaś zmiennym %on ich wyniki.

Zwyczajowo elementami, na których pracuje Maxima, są wyrażenia matematyczne składające się ze stałych i zmiennych. Oprócz nich obsługuje też łańcuchy tekstowe ujmowane w cudzysłów. Podstawowymi elementami Maximy są listy. Wszystkie struktury inne niż tablice czy tablice haszujące są listami. Lista jest zbiorem uporządkowanych elementów otoczonych nawiasami klamrowymi np. `[ a, b, 2]`.

Tablice są definiowane poleceniem `array`, któremu jako argument podaje się nazwę tablicy oraz maksymalne wartości indeksów. Liczbę wymiarów definiuje liczba podanych maksymalnych wartości indeksów. Trzeba pamiętać, że elementy tablic numerowane są od zera. Na przykład tablica `aa` o wymiarach 3×4 tworzy się poleceniem `array(aa, 2, 3)`. Tablica może mieć do pięciu wymiarów, a indeksy w poszczególnych wymiarach zaczynają się od zera i mogą się zmieniać do wartości zadeklarowanej przy tworzeniu tablicy. Tablicę wypełnia się, przypisując poszczególnym pozycjom w tablicy wartości, na przykład `aa[0,2]: %pi`. Przydatnym poleceniem jest `listarray(aa)`, które wyświetli zawartość tablicy. Oprócz wymienionych typów danych można też używać struktur, które stanowią zmienne zawierające różne pola.

Bardzo przydatną cechą Maximy jest możliwość generowania formuł w formacie bezpośrednio możliwym do użycia w językach programowania takich jak C, czy FORTRAN, a po niewielkich modyfikacjach można ich użyć w dowolnym języku programowania. Oczywiście Maxima może też robić wykresy i zapisywać je do pliku. Dla ułatwienia pracy z programem, warto też zapamiętać polecenie `batch("plik")`, które pozwala na wczytanie wcześniej opracowanego skryptu i wykonać obliczenia.

12.2.1. Wyrażenia i funkcje

Każde wyrażenie w systemie Maxima musi kończyć się średnikiem `;` lub znakiem dolara `$`. Średnik powoduje natychmiastową odpowiedź Maximy, natomiast znak dolara zapobiega wyświetleniu odpowiedzi. Wartości zmiennym przypisuje się za pomocą dwukropka `:`, na przykład `a: 5;`, `b: x^2-7*x`. Funkcje definiuje się operatorem dwukropek ze znakiem równości `:=`, na przykład `f(x):= x^2+cos(x)`.

W pakiecie Maxima zdefiniowano stałe używane w obliczeniach. Wybrane stałe matematyczne w pakiecie Maxima zostały zebrane w tabeli 12.1. Obliczenia przeprowadzane są symbolicznie, przy czym poleceniem `ev(a,b1,b2)` można wymusić wyznaczenie wartości wyrażeń symbolicznych dla arbitralnie podanych wartości zmiennych. Podobnie wyniki są podawane najczęściej jako wartości dokładne, co oznacza, że często zawierają ułamki zwykłe, wyrażenia niewymierne oraz stałe matematyczne.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 12.1. Podstawowe stałe w Maximie

Symbol	Znaczenie
%e	liczba e Eulera
%pi	liczba π
%phi	złota proporcja ($\phi = \frac{1+\sqrt{5}}{2}$)
%i	jednostka urojona ($i = \sqrt{-1}$)
%gamma	stała Eulera–Mascheroniego $\gamma = 0,5772156649015329\ldots$
inf	nieskończoność (∞)
minf	minus nieskończoność ($-\infty$)
infinity	nieskończoność w dziedzinie liczb zespolonych

Można dokonać przybliżenia wyniku do najbliższej wartości zmiennoprzecinkowej za pomocą polecenia `float(sqr(%pi))`. Aby zmienić precyzję wyświetlania wyniku można ustawić zmienną `fpprec` na wartość np. 60 `fpprec: 60`, by polecenie `bfloat(%pi)` zwróciło wartość z precyzją 60 cyfr znaczących. Można też wymusić przedstawienie wartości liczb zmiennoprzecinkowych w postaci wymiernej poleceniem `rationalize(a)`.

12.2.2. Różniczkowanie i całkowanie

W Maximie do symbolicznego różniczkowania i całkowania używa się poleceń, odpowiednio, `diff(f,x,n)` oraz `integrate(f, x)` lub `integrate(f, x, a, b)`. System usiłuje wykorzystać wiedzę, jaką posiada na temat różniczkowanych czy całkowanych funkcji, czy wyrażeń do wyznaczenia poprawnego wyniku.

Jeżeli podczas różniczkowania nie ma informacji na temat tego, która zmienna jest niezależna, to w wyniku zaznaczane to jest jako `del(x)`, co oznacza, że aby uzyskać pełne wyrażenie na pochodną, należy wykonać różniczkowanie zmiennej `x` względem zmiennej niezależnej. Zależność funkcji od zmiennych można zadeklarować poleceniem `depends([a,b],z)`, co oznacza, że funkcje `a` oraz `b` zależą od zmiennej `z`, która nie musi być niezależna.

Całkowanie przeprowadza się poleceniem `integrate(f,x)`, gdzie drugi argument jest zmienną, względem której odbywa się całkowanie. W ten sposób można wyznaczyć całkę nieoznaczoną. Do symbolicznego wyznaczania wartości całki oznaczonej służy to samo polecenie `integrate(f, x, a, b)` z dodatkowymi argumentami, które oznaczają odpowiednio dolną i górną granicę całkowania. Granice całkowania mogą być niewłaściwe tzn. wynosić $\pm\infty$, czyli `inf` lub `minf`.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Gdy nie udaje się wyznaczyć wartości całki symbolicznie, można posłużyć się metodami numerycznymi, by uzyskać wymagany wynik. W pakiecie zaimplementowano metodę Romberga (patrz podrozdział 4.4) wywoływaną poleceniem `romberg(f, x, a, b)`, na przykład `aa: romberg(cos(sin(x+1)), x, 0, 1);`.

12.2.3. Macierze

W systemie Maxima i pochodnych macierze definiuje się przy użyciu listy wierszy stanowiących argument polecenia `matrix`. Oznacza że, że przykładową macierz

$$a = \begin{bmatrix} 1-\lambda & 0 & 1 \\ 2 & 1-\lambda & 3 \\ 2 & 0 & 1-\lambda \end{bmatrix}$$

można wprowadzić za pomocą polecenia `a:matrix([1-lambda, 0, 1],[2, 1-lambda, 3],[2, 0, 1-lambda]);`. Macierze o wymiarach $n \times m$ można też wprowadzać interaktywnie za pomocą polecenia `entermatrix(n,m)`.

Wyznacznik macierzy `a` można wyznaczyć poleceniem `determinant(a)`, które zwróci w wyniku wyrażenie

$$|a| = (1-\lambda)^3 - 2\,(1-\lambda)$$

Wartości i wektory własne macierzy można wyznaczyć za pomocą polecenia odpowiednio `eigenvalues(a)` oraz `eigenvectors(a)`, które dadzą wyniki następujące:

— polecenie `eigenvalues(a);`

$$\left[\left[-\lambda-\sqrt{2}+1, -\lambda+\sqrt{2}+1, 1-\lambda\right], [1,1,1]\right],$$

— polecenie `eigenvectors(a);` oprócz wartości własnych zwróci wektory, które dla rozważanego przykładu wyniosą

$$\left[1, 3-\sqrt{2}, -\sqrt{2}\right],\ \left[1, \sqrt{2}+3, \sqrt{2}\right],\ [0,1,0].$$

Kwadratowe macierze można odwracać poleceniem `invert(a)`. Transpozycję macierzy wyznacza się za pomocą polecenia `transpose(a)`. Mnożenie macierzy realizuje się operatorem kropki `.`, który dla wektorów o elementach rzeczywistych realizuje iloczyn skalarny. Należy pamiętać, że za pomocą tradycyjnych operatorów `+,-,*,/,^` wykonuje się operacje element po elemencie i oznaczają one odpowiednio dodawanie, odejmowanie, mnożenie, dzielenie i potęgowanie (element po elemencie). W terminologii MATLAB-a takie operatory są nazywane operatorami tablicowymi. Potęgowanie macierzowe można wykonać za pomocą operatora `^^`, na przykład `a^^2`. Nic nie stoi na przeszkodzie, by w macierzy umieścić wyrażenia lub funkcje oraz użyć funkcji do różniczkowania i całkowania w stosunku do macierzy.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Aby zilustrować to stwierdzenie, zostanie zaprezentowany przykład wyznaczenia macierzy sztywności elementu prętowego. Postępując zgodnie z tokiem zaprezentowanym w rozdziale 11, można wyznaczyć postać tej macierzy. Poniższy listing przedstawia skrypt pozwalający wyznaczyć macierz sztywności.

```
  xi: x/l;
2 N: matrix([1-xi,xi]);
  B: diff(N,x);
4 K: E*A*integrate(transpose(B).B,x,0,l);
```

Wynikiem obliczeń jest macierz w postaci

$$K = \begin{bmatrix} \frac{A\,E}{l} & -\frac{A\,E}{l} \\ -\frac{A\,E}{l} & \frac{A\,E}{l} \end{bmatrix}$$

co jak łatwo sprawdzić, jest prawidłowym wynikiem, zgodnym z obliczeniami z rozdziału 11. Można jedynie dokonać pewnych uproszczeń, wyłączając wspólny czynnik przed macierz.

12.2.4. Równania i układy równań

Aby rozwiązać równanie lub układ równań algebraicznych liniowych czy nieliniowych, używa się polecenia `solve(f,z)`. W poleceniu tym argumentami są: równanie lub lista równań oraz niewiadoma lub lista niewiadomych, których wartości należy wyznaczyć. Na przykład rozwiązanie równania

$$x^2 - 4 = 0$$

za pomocą pakietu Maxima może wyglądać tak: `solve(x^2-4=0,x);`. Przy przyrównywaniu do zera można pominąć znak równości i prawą stronę, upraszczając zapis do `solve(x^2-4,x)`.

Aby zilustrować dotychczas omówione zagadnienia, zostanie przedstawiony przykład wyznaczenia wartości współczynników w funkcji interpolującej rozkład przemieszczeń w zginanej belce zgodnie z równaniami (11.112).

```
  fKsztalt: a_1+a_2*x+a_3*x^2+a_4*x^3;
2 r1: subst(0,x,fKsztalt);
  r2: subst(l,x,fKsztalt);
4 poch: -diff(fKsztalt,x);
  r3: subst(0,x,poch);
6 r4: subst(l,x,poch);
  solve([r1=v_1,r2=v_2,r3=theta_1,r4=theta_2],[a_1,a_2,a_3,a_4]);
```

Pierwsza linia zawiera ogólną postać funkcji interpolującej rozkład przemieszczeń zgodnie z równaniem (11.110). Linie druga i trzecia podstawiają za zmienną x lewy (linia 2) i prawy brzeg (linia 3) elementu i wyznaczają lewe strony równań.

Linia czwarta służy do wyznaczenia pochodnej względem zmiennej x, zgodnie z zależnością (11.111), co opisuje zmiany kąta obrotu przekrojów belki wzdłuż jej

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

długości. Z kolei linie 5 i 6 budują lewe strony równań uwzględniających warunki brzegowe związane z kątami obrotu przekrojów. W linii szóstej wykorzystano polecenie `solve` do rozwiązania układu równań ze względu na nieznane współczynniki wielomianu interpolującego.

12.2.5. Równania różniczkowe

Rozwiązywanie równań różniczkowych zazwyczaj jest wyzwaniem. Tylko pewne klasy tych równań można bez problemu rozwiązać, inne stanowią poważny problem. Także w tym zakresie Maxima może wesprzeć inżyniera w jego pracy.

W przypadku symbolicznego rozwiązywania równań różniczkowych zwyczajnych pierwszego i drugiego rzędu można je rozwiązywać przez bezpośrednie poszukiwanie funkcji spełniających równanie lub wykorzystać transformację Laplace'a, by problem analizy matematycznej zapisać w postaci równań algebraicznych, a następnie dokonać transformacji odwrotnej.

Powyższe stwierdzenia zostaną zilustrowane przykładem polegającym na rozwiązaniu równania różniczkowego z przykładu w podrozdziale 6.2.4 (patrz strona 129), które ma następującą postać

$$5\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \frac{\mathrm{d}x}{\mathrm{d}t} + 8x = 0, \quad x(0) = 0,1 \quad \left.\frac{\mathrm{d}x}{\mathrm{d}t}\right|_{t=0} = 0$$

Tradycyjne rozwiązanie powyższego równania przedstawiono w poniższym listingu.

```
1 equation: 5*'diff(x,t,2)+'diff(x,t)+8*x = 0;
  solution: ode2(equation,x,t);
3 rozw: ic2(solution,t=0,x=0.1,'diff(x,t)=0);
```

W pierwszej linii zostaje zdefiniowane równanie i przypisane zmiennej `equation`. Następnie wywoływana jest funkcja `ode2`, która służy do rozwiązania równania różniczkowego, a następnie uwzględniane są warunki początkowe za pomocą funkcji `ic2`, zaś samo rozwiązanie przypisane zostaje zmiennej `rozw`.

Rozwiązanie tego samego równania za pomocą transformacji Laplace'a można wykonać z wykorzystaniem poniższej sekwencji poleceń.

```
1 equation: 5*'diff(x(t),t,2)+'diff(x(t),t)+8*x(t) = 0;
  atvalue(x(t),t=0,0.1);
3 atvalue('diff(x(t),t),t=0,0);
  lapOde: laplace(equation,t,s);
5 sol: solve(lapOde, 'laplace(x(t), t, s));
  map(lambda( [eq], ilt(eq, s, t)), sol);
```

Podobnie, jak poprzednio, w pierwszej linii jest definiowane równanie różniczkowe. Ważne jest, by zastosować polecenie opóźnionego różniczkowania `'diff` do zapisu odpowiednich pochodnych. Linie druga i trzecia pozwalają uwzględnić warunki początkowe za pomocą polecenia `atvalue`, które może też służyć do nakładania warunków brzegowych lub do określania wartości, którą rozwiązanie ma przyjąć dla jakiegoś argumentu. Linia czwarta realizuje transformację Laplace'a równania różniczkowego.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zmienna x(t) zostaje przetransformowana do laplace(x(t), t, s). W linii piątej rozwiązywane jest równanie algebraiczne uzyskane po transformacji Laplace'a ze względu na transformowaną zmienną. Ostatnia linia powoduje wykonanie odwrotnej transformacji Laplace'a, by powrócić do dziedziny czasu.

Znacznie prościej można uzyskać takie rozwiązanie z wykorzystaniem transformacji Laplace'a, wykorzystując polecenie Maximy desolve. Poniżej zamieszczono przykład zastosowania tego polecenia.

```
  equation: 5*'diff(x(t),t,2)+'diff(x(t),t)+8*x(t) = 0;
2 atvalue(x(t),t=0,0.1);
  atvalue('diff(x(t),t),t=0,0);
4 solution: desolve(equation,x(t));
```

Pierwsze trzy linie zostały już omówione przy poprzednim listingu. Ostatnia wywołuje polecenie dokonujące transformacji Laplace'a równania różniczkowego, rozwiązuje równanie algebraiczne, a następnie dokonuje odwrotnej transformacji.

12.2.6. Pozostałe możliwości

Maxima potrafi nie tylko rozwiązywać równania, całkować czy różniczkować. Za jej pomocą można też rozwijać funkcje w szeregi, rozwiązywać zadania z teorii liczb, czy przeprowadzać proste analizy statystyczne, a nawet przeliczać i upraszczać jednostki. A jest to tylko wąski wycinek zagadnień, do których można zastosować ten pakiet oprogramowania. Otwartość i dostępność dużej liczby dodatkowych pakietów to zalety, które trudno przecenić. W dalszej części tego podrozdziału zostanie przedstawiony przykład nieco bardziej złożonych obliczeń, które posłużą do przybliżonego rozwiązania równania różniczkowego metodą Galerkina.

Przykład rozwiązania równania różniczkowego metodą Galerkina. Należy rozwiązać poniższe równanie różniczkowe zwyczajne drugiego rzędu

$$\frac{\mathrm{d}^2T}{\mathrm{d}x^2} + \frac{\mathrm{d}T}{\mathrm{d}x} + 20x^3 = 0$$

dla $x \in \langle 0, 1\rangle$ z warunkami brzegowymi $x(0) = 0$, $x(1) = 0$.

Do przybliżonego rozwiązania problemu zostały arbitralnie wybrane dwie funkcje próbne spełniające warunki brzegowe. Dla zademonstrowania metody z rozmysłem wybrano takie funkcje próbne, które nie stanowią dokładnego rozwiązania analizowanego równania różniczkowego.

$$N_1 = x\left(1 - x^2\right), \quad N_2 = x\left(1 - x^4\right)$$

Następnie wyznaczono residua i z warunku zerowania całki z ważonego residuum wyliczono współczynniki przybliżonego rozwiązania. Dokładniej postępowanie zostało opisane w podrozdziale 11.4.2, zaś cały tok obliczeń wraz z poleceniami generującymi odpowiedni wykres został przedstawiony na poniższym listingu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

```
/* granice obszaru analizy */
gd: 0;
gg: 1;

/*  funkcje probne */
N_1: x*(1-x^2);
N_2: x*(1-x^4);

/* rozwiazanie w postaci */
Tp: a_1*N_1+a_2*N_2;

residua: diff(Tp,x,2)+diff(Tp,x,1)+20*x^3;

/* calki reziduow wazonych */
r1: integrate(N_1*residua,x,gd,gg);
r2: integrate(N_2*residua,x,gd,gg);

/* rozwiazanie ukladu rownan */
rozw: solve([r1=0,r2=0],[a_1,a_2]);

/* parametry */
a1: rhs(rozw[1][1]);
a2: rhs(rozw[1][2]);

/* rozw. dokladne */
equation: 'diff(T,x,2)+'diff(T,x,1)=-20*x^3;
solution: ode2(equation,T,x);

/* problem brzegowy - funkcja bc2 naklada war. brzegowe */
Rozw:bc2(solution,x=0,T=0,x=1,T=0);

/* rysowanie obu rozwiazan */
plot2d([a1*N_1+a2*N_2, rhs(Rozw)],[x,gd,gg],[ylabel, "T"],grid2d,
  [legend,"Tp(x)","T(x)"]);
```

W liniach od 1 do 24 odbywa się rozwiązanie przybliżone, zaś od 25 linii przedstawiono sposób ścisłego rozwiązania równania różniczkowego. Linie 33 oraz 34 zawierają polecenie pozwalające graficznie ocenić dokładność rozwiązania.

Po uruchomieniu opisanego skryptu, sesja Maximy będzie przyjmować wygląd zbliżony do poniższego:

```
(%i1) batch("galerkin.maxima");

read and interpret file: #p/galerkin.maxima
(%i2) gd:0
(%o2)                                  0
(%i3) gg:1
(%o3)                                  1
(%i4) N_1:x*(1-x^2)
                                         2
(%o4)                             x (1 - x )
```

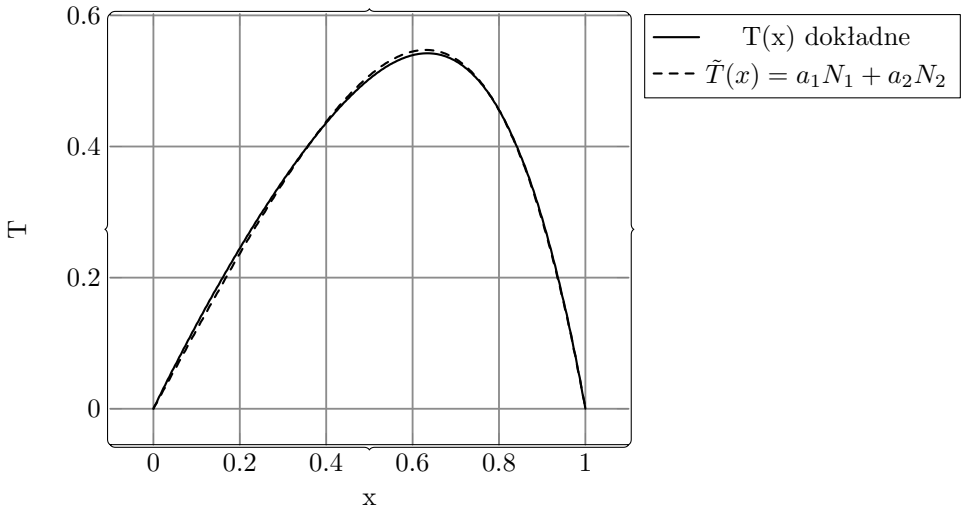
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

```
(%i5) N_2:x*(1-x^4)
                                        4
(%o5)                                x (1 - x )
(%i6) Tp:a_2*N_2+a_1*N_1
                                      4                 2
(%o6)                    a_2 x (1 - x ) + a_1 x (1 - x )
(%i7) residua:20*x^3+diff(Tp,x,1)+diff(Tp,x,2)
               4              4            3       3          2
(%o7) (- 4 a_2 x ) + a_2 (1 - x ) - 20 a_2 x  + 20 x  - 2 a_1 x
             2
+ a_1 (1 - x ) - 6 a_1 x
(%i8) r1:integrate(N_1*residua,x,gd,gg)
                            (- 960) + 672 a_1 + 925 a_2
(%o8)                     - --------------------------
                                       840
(%i9) r2:integrate(N_2*residua,x,gd,gg)
                            (- 896) + 597 a_1 + 896 a_2
(%o9)                     - --------------------------
                                       504
(%i10) rozw:solve([r1 = 0,r2 = 0],[a_1,a_2])
                                 31360        9664
(%o10)                   [[a_1 = -----, a_2 = -----]]
                                 49887        16629
(%i11) a1:rhs(rozw[1][1])
                                    31360
(%o11)                              -----
                                    49887
(%i12) a2:rhs(rozw[1][2])
                                    9664
(%o12)                              -----
                                    16629
(%i13) equation:'diff(T,x,1)+'diff(T,x,2) = -20*x^3
                                2
                               d T   dT         3
(%o13)                         --- + -- = - 20 x
                                 2   dx
                               dx
(%i14) solution:ode2(equation,T,x)
                      - x      4       3       2
(%o14)     T = %k2 %e    - 5 x  + 20 x  - 60 x  + 120 x + %k1 - 120
(%i15) Rozw:bc2(solution,x = 0,T = 0,x = 1,T = 0)
                1 - x
           75 %e           4       3       2           45 %e - 120
(%o15) T = ---------- - 5 x  + 20 x  - 60 x  + 120 x + ----------
             %e - 1                                      %e - 1
```

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

```
        - 120
(%i16) plot2d([a2*N_2+a1*N_1,rhs(Rozw)],[x,gd,gg],[ylabel,"T"],grid2d,
             [legend,"Tp(x)","T(x)"])
```

Skrypt wyświetla porównanie dwóch rozwiązań – przybliżonego i dokładnego. Graficzny wynik działania skryptu przedstawiono na rysunku 12.1. W ten sposób można przygotowywać skrypty pozwalające rozwiązywać problemy napotykane w życiu codziennym studenta czy inżyniera.



Rysunek 12.1. Porównanie rozwiązania dokładnego i przybliżonego metodą Galerkina

Należy sobie zdawać sprawę, że podstawową formą pracy z systemami algebry komputerowej jest interaktywna sesja. Mimo wszystko należy stale kontrolować jakość otrzymywanych wyników i na bieżąco reagować na pojawiające się problemy. Jedynie dobrze poznane problemy, które operator może uznać za standardowe, można automatyzować.

12.2.7. Wybrane funkcje

Dotychczas opisane polecenia i operatory pozwalają na stawianie pierwszych kroków. Poniżej przedstawiono wybór poleceń, które mogą być przydatne dalszej pracy z tym programem. Jeżeli krótki opis nie jest wystarczający, warto zapoznać się z pomocą dostępną w samym pakiecie Maxima.

`allroots(a)`:
wyznacza wszystkie pierwiastki wielomianu a oraz wyświetla je w postaci numerycznej

`append(a,b)`:
łączy listy a i b, zwracając pojedynczą listę

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

batch("plik"):
uruchamia skrypt Maximy zapisany w pliku *plik*

coeff(a,b,c):
zwraca współczynnik stojący przy b podniesionym do potęgi c w wyrażeniu a

concat(a,b):
łączy łańcuchy a i b, zwracając tekst stanowiący ich połączenie

cons(a,b):
dodaje na początku listy b element a

demoivre(a):
przekształca liczby zespolone z postaci wykładniczej do trygonometrycznej

denom(a):
zwraca mianownik a

depends(a,b):
deklaracja, że a jest funkcją zmiennej b, przydatne przy zapisie równań różniczkowych

desolve(a,b):
rozwiązuje układ liniowych równań różniczkowych zwyczajnych a względem zmiennej b, z wykorzystaniem transformacji Laplace'a

determinant(a):
zwraca wyznacznik macierzy a

diff(a,b1,c1,..,bn,cn):
zwraca c_i-tą pochodną cząstkową względem zmiennych b_i, `diff(a,b,1)` można skrócić do `diff(a,b)`, `'diff(...)` służy do zapisu pochodnych w równaniach różniczkowych

eigenvalues(a):
zwraca wartości własne macierzy a oraz ich krotności

eigenvectors(a):
jak `eigenvalues` plus lista wektorów własnych

entermatrix(a,b):
interaktywne wprowadzanie wartości elementów macierzy $a \times b$

ev(a,b1,b2,...,bn):
wykonuje działanie a, uwzględniając warunki b_i; w szczególności b_i mogą być równaniami, listami równań, przypisaniami, które `ev` wstawia do a; b_i może też być tekstem: `numer` (wynik zwracany jest w formacie zmiennoprzecinkowym), `detout` (odwracanie macierzy wykonywane z wyznacznikiem wyłączonym przed macierz) oraz `diff` (wszystkie różniczkowania są wykonywane, czyli `'diff` jest zamieniane na `diff`)

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

expand(a):
rozwija wyrażenie algebraiczne a, np. mnoży wyrażenia w nawiasach i je redukuje

exponentialize(a):
zapisuje funkcje trygonometryczne w postaci wykładniczej

factor(a):
rozkłada na czynniki pierwsze liczbę a

fortran(a):
zwraca wyrażenie a w formacie języka Fortran, użyteczne do przenoszenia do MATLAB-a/GNU Octave

freeof(a,b):
zwraca wartość logiczną: prawda, gdy zmienna a nie jest częścią wyrażenia b

grind(a):
wyświetla zmienną lub funkcję w zwięzłym formacie, przydatnym do tworzenia plików wsadowych z wyrażeniami Maximy

horner(a,b):
zwraca wyrażenie a przekształcone do postaci analogicznej do schematu Hornera, używając b jako zmiennej niezależnej

ident(a):
zwraca macierz jednostkową o rozmiarze $a \times a$

imagpart(a):
zwraca część urojoną a

integrate(a,b):
wyznacza całkę nieoznaczoną z a względem zmiennej b

integrate(a,b,c,d):
wyznacza całkę oznaczoną z a względem b w granicach od c do d, granice mogą być nieskończone `inf` lub `minf`

invert(a):
oblicza macierz odwrotną do a

kill(a):
usuwa z przestrzeni roboczej zmienną a wraz ze wszystkimi przypisaniami i właściwościami

limit(a,b,c):
daje granicę wyrażenia a, gdy zmienna b zmierza do c

lhs(a):
zwraca lewą stronę równania a

loadfile(a):
wczytuje plik a z bieżącego katalogu, plik musi być w odpowiednim formacie – na przykład wygenerowany poleceniem `save`

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

makelist(a,b,c,d):
tworzy listę na podstawie wyrażenia a zależnego od b, gdzie b zmienia się od c do d, np. `makelist(7/z,z,1,5)`

map(a,b):
wykonuje funkcję a na wyrażeniach składowych b

matrix(a1,a2,...,an):
tworzy macierz $n \times m$, której wiersze są listami `ai=[b1, b2, ...,bm]`

num(a):
zwraca licznik a

ode2(a,b,c):
rozwiązuje równania różniczkowe zwyczajne a pierwszego i drugiego rzędu względem zmiennej b będącej funkcją zmiennej c

part(a,b1,b2,...,bn):
zwraca b_i-tą część wyrażenia a, np. `part(w-y+2,2)` daje y, części są numerowane tak jak w wyrażeniu wyświetlonym przez Maximę

ratsimp(a):
upraszcza wielomian a i zwraca wynik jako iloraz dwóch wielomianów

realpart(a):
zwraca część rzeczywistą a

rhs(a):
zwraca prawą stronę równania a

save(a,b1,b2,..., bn):
zapisuje w bieżącym katalogu plik zawierający zmienne, funkcje i tablice b_i i nadaje mu nazwę a, $b1$ może przyjąć wartość `all`, wtedy zapisywane są wszystkie dane z przestrzeni roboczej

solve(a,b):
rozwiązuje algebraiczne równanie a ze względu na niewiadomą b; a i b mogą być listami i wtedy rozwiązywany jest układ równań

subst(a,b,c):
podstawia a za b w wyrażeniu c

taylor(a,b,c,d):
zwraca rozwinięcie a w szereg Taylora

transpose(a):
wykonuje transpozycję macierzy a

trigexpand(a):
upraszcza wyrażenia zawierające funkcje trygonometryczne z wykorzystaniem wzorów na sumy i różnice kątów

trigreduce(a):
upraszcza wyrażenia trygonometryczne z wykorzystaniem tożsamości

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

trygonometrycznych tak, by zamienić iloczyny i potęgi funkcji sinus i cosinus na sumy wyrażeń z pojedynczym sinusem lub kosinusem; na przykład `trigreduce(sin(x)^2)` daje `(1 - cos(2x))/2`

`trigsimp(a)`:
upraszcza wyrażenia trygonometryczne zawierające funkcje tangens, cotangens i inne, zamieniając je na odpowiedniki z sinusami i kosinusami; używa też tożsamości $\sin^2 x + \cos^2 x = 1$

12.3. SageMath

Bardzo ciekawą propozycją pakietu obliczeniowego o bardzo uniwersalnym zastosowaniu jest SageMath. Napisany w języku Python pakiet pozawala na prowadzenie obliczeń numerycznych oraz symbolicznych. Tak naprawdę jest to platforma, która integruje różne narzędzia matematyczne. Wykorzystuje między innymi Maximę, GNU Octave, środowisko obliczeń statystycznych R oraz moduły obliczeniowe Pythona. W wersji 7.3 SageMath wykorzystywało ponad 90 różnych otwartych pakietów oprogramowania.

Pakiet jest przeznaczony dla matematyków, więc obliczenia prowadzone są dokładnie. Aby uzyskać wyniki numeryczne (przybliżone), konieczne jest stosowanie dodatkowych zabiegów. Dzięki temu można ustrzec się wielu błędów, które w innych systemach obliczeniowych pojawiają się niezależnie od woli użytkownika.

Bardzo ciekawym elementem tego pakietu jest `sagetex`, który pozwala wykonywać obliczenia bezpośrednio w dokumentach składanych za pomocą LaTeX-a. Daje to duże możliwości przy tworzeniu dokumentów typu sprawozdania, raporty czy opracowywanie wyników z zajęć laboratoryjnych.

Podstawową cechą języka SageMath jest jego obiektowość. Wszystkie elementy wprowadzane do obliczeń są obiektami. Zatem by dokonywać różnych konwersji, modyfikacji czy wyświetlać wynik z zadaną precyzją, wykorzystuje się metody obiektów.

Mimo że do wielu symbolicznych obliczeń SageMath używa Maximy, to istnieją znaczne różnice w składni obu pakietów. Jedną z istotniejszych jest sposób przypisywania zmiennym wartości. Używa się do tego celu znaku równości =, natomiast przy definiowaniu równań używa się znaku ==. W SageMath obowiązuje konwencja stosowana w języku Python związana z kończeniem linii znakiem końca linii, zaś bloki programu wydziela się wcięciami.

Należy pamiętać, że SageMath to bardzo rozbudowany pakiet, pozwalający na prowadzenie bardzo restrykcyjnych obliczeń z uwzględnieniem szczegółów, takich jak przestrzenie, grupy czy ciała, do których poszczególne obiekty należą. W związku z tym niniejszy opis należy traktować jako bardzo zgrubne przedstawienie wąskiego wycinka możliwości pakietu.

12.3.1. Wyrażenia i funkcje

Niektóre stałe zdefiniowane w SageMath zawiera tabela 12.2. Chociaż omawiana platforma stanowi zbiór różnych pakietów i bibliotek matematycznych, to stałe defi-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

niuje się w spójny sposób. Wewnętrzne mechanizmy SageMath dokonują niezbędnych translacji.

Tabela 12.2. Wybrane stałe dostępne w SageMath

Symbol	Znaczenie
`e`	liczba e Eulera
`pi`	liczba π
`golden_ratio`	złota proporcja ($\phi = \frac{1+\sqrt{5}}{2}$)
`I`	jednostka urojona ($i = \sqrt{-1}$)
`euler_gamma`	stała Eulera–Mascheroniego $\gamma = 0,5772156649015329\ldots$
`infinity`	nieskończoność (∞)
`-infinity`	minus nieskończoność ($-\infty$)

12.3.2. Różniczkowanie i całkowanie

W SageMath różniczkowanie realizuje się poleceniem `derivative`, zaś całkowanie poleceniem `integral`. W pakiecie zdefiniowano aliasy, które mają identyczną składnię, aczkolwiek nieco różne nazwy. Do całkowania takie aliasy to `integral` oraz `integrate`. Składnia polecenia do całkowania jest następująca:

`integral(wyr, v=None, a=None, b=None, algorithm=None, hold=False)`.

Pierwszy argument polecenia jest obowiązkowy i oznacza wyrażenie, które należy scałkować. Drugi argument to zmienna niezależna, względem której ma przebiegać całkowanie. Może to być kilka zmiennych lub też zmienna z granicami przedziału całkowania np. `(x,-2,2)`. Następne dwa argumenty to granice całkowania. Jeżeli nie zostaną podane granice całkowania, wyznaczana jest całka nieoznaczona, w przeciwnym przypadku całka oznaczona. Można wybrać algorytm, którego SageMath użyje do całkowania symbolicznego. Dostępne są następujące możliwości `'maxima'`, `'sympy'`, `'mathematica_free'`, `'fricas'`. Domyślnym algorytmem jest Maxima. Sympy oznacza użycie algorytmu z pakietu Sympy – biblioteki języka Python do obliczeń symbolicznych. Możliwe jest użycie algorytmu używanego przez Mathematicę. Aby taki algorytm zadziałał, konieczne jest połączenie z internetem, ponieważ obliczenia przeprowadzane są zdalnie pod adresem `http://integrals.wolfram.com/`. Co ważne, takie użycie algorytmu z programu Mathematica nie wymaga żadnych opłat licencyjnych[2]. Ostatnia możliwość dotyczy użycia pakietu `FriCAS` do całkowania. Wybór tego algorytmu wiąże się z zainstalowaniem w systemie tego pakietu.

[2] Stan na 2017 rok.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Podobnie jak w Maximie, niekiedy należy podać spełniane przez funkcję podcałkową dodatkowe warunki, z których wynikają pewne założenia dla algorytmów całkowania. Można zrobić to interaktywnie, odpowiadając na pytania środowiska lub wykorzystując polecenie `assume`, które pozwala podawać dodatkowe warunki rozwiązania, np. `assume(a>0)`.

Oprócz całkowania symbolicznego jest też możliwe wykonywanie obliczeń numerycznych. Służy do tego polecenie `numerical_integral(func, a, b=None, algorithm ='qag', max_points=87, params=[], eps_abs=1e-06, eps_rel=1e-06, rule=6)`. Przy numerycznym wyznaczaniu wartości całki oznaczonej można wpływać na użyty algorytm oraz wymaganą dokładność obliczeń. Dokładność można podawać jako wartość bezwzględną lub dopuszczalny błąd względny. Granice `a` oraz `b` mogą być niewłaściwe, w postaci `Infinity` lub `-Infinity`. Dostępne algorytmy całkowania to `'qag'` – adaptacyjna metoda całkowania, `'qags'` – metoda adaptacyjna z detekcją osobliwości oraz `'qng'` – kwadratura Gaussa–Kronroda (patrz podrozdział 4.7.1). Maksymalna liczba węzłów kwadratury jest zadawana parametrem `max_points`. Ostatni parametr o nazwie `rule` decyduje o użytej kwadraturze Gaussa–Kronroda, i tak

Parametr	Liczba węzłów	Parametr	Liczba węzłów
`rule=1`	15	`rule=4`	41
`rule=2`	21	`rule=5`	51
`rule=3`	31	`rule=6`	61

Większe wartości parametru `rule` dają dokładniejsze wyniki dla gładkich funkcji, mniejsze są dokładniejsze dla funkcji z nieciągłościami.

12.3.3. Macierze

W SageMath macierze i wektory definiuje się za pomocą poleceń, które tworzą odpowiednie obiekty. Macierz można zdefiniować poleceniem `matrix`, na przykład aby utworzyć macierz

$$M = \begin{bmatrix} 1-\omega^2 & 0 & 1 \\ 2 & 1-\omega^2 & 3 \\ 2 & 0 & 1-\omega^2 \end{bmatrix}$$

najpierw należy zadeklarować identyfikator `omega` jako zmienną symboliczną, a następnie utworzyć odpowiednią macierz. Ten sposób postępowania zapisano w postaci programu na następnej stronie.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

W pierwszej linii poniższego programu deklaruje się zmienną symboliczną. Następna linia tworzy macierz kwadratową z trzema kolumnami i wierszami.

```
var('omega')
M= matrix(3,3,[[1-omega^2, 0, 1],[2, 1-omega^2, 3],[2, 0, 1-omega^2]])
```

Warto zwrócić uwagę na zasadniczą różnicę między SageMath a Maximą: w SageMath macierz jest listą list zawierających wiersze macierzy. Ponieważ macierz `M` jest obiektem, wszystkie operacje na tej macierzy realizuje się przez wywoływanie odpowiednich metod obiektu `M`. Służy do tego operator kropki. Na przykład wyznacznik macierzy `M` obliczyć można poleceniem `M.det()` lub `M.determinant()`. Natomiast transpozycję macierzy `M` uzyskuje się przez wywołanie metody `transpose`, czyli należy wpisać `M.transpose()`.

Aby się zorientować, jakie metody udostępnia dany obiekt, trzeba wpisać nazwę obiektu, kropkę i nacisnąć klawisz `Tab`. Zatem listę metod dotyczących macierzy `M` można uzyskać wpisując `M.`, a następnie nacisnąć klawisz `Tab`. Nazwy metod zazwyczaj są na tyle opisowe, że można się zorientować czego dotyczą. Pomoc można uzyskać, wpisując nazwę metody i na końcu dodając znak zapytania, np. `M.transpose?`. Niekiedy lista dostępnych metod może być dosyć długa, warto ją skrócić. Gdy znana jest początkowa część nazwy metody, można ją wpisać po znaku kropki, a naciśnięcie klawisza tabulacji wyświetli listę metod, których nazwy rozpoczynają się w ten sposób. Poniżej przedstawiono to na przykładzie fragmentu interaktywnej sesji w pakiecie SageMath. Dzięki zaimplementowanemu w systemie autouzupełnianiu nie trzeba pamiętać nazw wszystkich metod, które można zastosować do danego obiektu.

```
sage: M.d
M.db                          M.density
M.decomposition               M.derivative
M.decomposition_of_subspace   M.det
M.delete_columns              M.determinant
M.delete_rows                 M.diagonal
M.denominator                 M.dict
M.dense_columns               M.dimensions
M.dense_matrix                M.dump
M.dense_rows                  M.dumps
```

12.3.4. Równania i układy równań

Jak już wspomniano, równania w SageMath deklaruje się z wykorzystaniem operatora == stojącego między wyrażeniami określającymi lewą i prawą stronę równania. Przed zdefiniowaniem równania należy zadeklarować zmienne symboliczne za pomocą polecenia `var`, na przykład `var('x,a,b')`. Rozwiązanie równania realizuje się za pomocą funkcji `solve`, której podaje się jako argument równanie oraz zmienną, ze względu na którą należy równanie rozwiązać. Na przykład

`rozw = solve(2*x^2-b^2 == 7, x)`

rozwiąże równanie $2x^2 - b^2 = 7$, wyznaczając wartość niewiadomej x. Rozwiązania są zwracane w formie listy równań

`[x == -sqrt(1/2*b^2 + 7/2), x == sqrt(1/2*b^2 + 7/2)]`.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Do odpowiednich elementów tej listy można się odwoływać przez indeksy w nawiasach kwadratowych `rozw[0]`, natomiast lewą i prawą stronę równania można uzyskać metodami `lhs()` oraz `rhs()`. Na przykład `rozw[0].rhs()` zwróci

`-sqrt(1/2*b^2 + 7/2)`.

W ten sposób można w skryptach wykorzystywać wyniki funkcji `solve`.

Przy rozwiązywaniu równań należy pamiętać o tym, że w języku Python, w którym jest pisany SageMath, operator ^ realizuje binarną funkcję XOR. SageMath przedefiniowuje znaczenie tego operatora. Jednak przy wykorzystywaniu efektów działania SageMath warto pamiętać, że bezpieczniej jest używać operatora ** do wprowadzania potęgowania.

Przykład rozwiązywania równań. Równanie stanu gazu doskonałego Clapeyrona rozszerzone przez Van der Waalsa ma postać

$$\left(p + \frac{a}{V_m^2}\right)(V_m - b) = RT \tag{12.1}$$

gdzie a jest stałą charakterystyczną dla danego gazu, uwzględniającą oddziaływanie między cząsteczkami gazu; b to stała charakterystyczna dla danego gazu, uwzględniająca skończone rozmiary cząsteczek, ma wymiar objętości, przez co uznawana jest za objętość mola cząsteczek gazu; p jest ciśnieniem; V_m jest objętością molową; zaś T to temperatura bezwzględna; natomiast R to uniwersalna stała gazowa.

Znając wartości $R = 0{,}08$, $a = 3{,}592$, $b = 0{,}04267$, $p = 10$ oraz $T = 300$ i zakładając, że jednostki tworzą spójny układ, wyznacz objętość molową V_m.

Rozwiązanie tego problemu można zrealizować numerycznie lub symbolicznie. Symboliczne rozwiązanie za pomocą SageMath może wyglądać tak, jak zapisano w poniższym listingu. Po zadeklarowaniu zmiennych jako symboliczne, w drugiej linii rozwiązano równanie, które zostało zdefiniowane jako argument polecenia `solve`. Do zapisu potęgowania użyto operatora stosowanego w języku Python.

```
var('p a V_m b R T')
a=solve([(p+a/V_m**2)*(V_m-b)==R*T],[V_m])
wyn=simplify(a[2])
latex(wyn)
```

Ponieważ równanie ma trzy rozwiązania, zostały one zwrócone jako lista rozwiązań. Jak wcześniej wspomniano, pojedyncze elementy rozwiązania można uzyskać, wpisując `a[0]`, `a[1]` oraz `a[2]`. Dwa z rozwiązań są zespolone, natomiast trzecie rozwiązanie `a[2]`, po pewnych uproszczeniach, zostało przedstawione poniżej

$$V_m = \left(\frac{1}{45\cdot 10^{11}}\sqrt{2\,292\,362\,572\,067\,434\,194\,929}\sqrt{6} + \frac{10\,833\,108\,114\,582\,163}{27\cdot 10^{15}}\right)^{\frac{1}{3}} + \\ + \frac{48\,890\,367\,289}{9\cdot 10^{10}\left(\frac{1}{45\cdot 10^{11}}\sqrt{2\,292\,362\,572\,067\,434\,194\,929}\sqrt{6} + \frac{10\,833\,108\,114\,582\,163}{27\cdot 10^{15}}\right)^{\frac{1}{3}}} + \\ + \frac{244\,267}{3\cdot 10^{5}} \tag{12.2}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po rozwiązaniu symbolicznym należy wstawić odpowiednie wartości zmiennych. Eksport do systemu składu tekstu LaTeX realizuje się poleceniem `latex(a[2])`. Próbę uproszczenia można podjąć, używając funkcji `simplify`, na przykład `simplify(a[2])`.

Jak widać z równania (12.2), elegancja i zwięzłość zostały złożone na ołtarzu dokładności. Po drobnych ręcznych modyfikacjach wyniku można go zapisać w zwięzłej postaci jako

$$V_m = 2{,}2886 \tag{12.3}$$

Po raz kolejny należy wspomnieć, że tego typu zadania należy rozwiązywać w trybie interaktywnym.

12.3.5. Równania różniczkowe

Przy pracy w SageMath, rozwiązywanie równań różniczkowych może się odbywać za pomocą Maximy. Między innymi, z tego powodu przebiega ono podobnie jak w Maximie, chociaż występują pewne różnice.

Na początku należy zdefiniować równanie, ale nie trzeba już, tak jak w Maximie, używać operatora opóźnionego różniczkowania, tylko użyć zwykłego polecenia `diff`. Warunki początkowe podaje się jako trzeci argument polecenia `desolve` lub też używa się składni `ics=[x0, y0,dy(x0)/dx, ...]`. Można to prześledzić na poniższych przykładach. Rozwiązanie równania różniczkowego rzędu pierwszego zamieszczono na poniższym listingu.

```
  y = function('y')(x)
2 de = diff(y,x) - x*y -2
  solution = desolve(de, y, ics=[0,1])
```

Równanie różniczkowe rzędu drugiego można rozwiązać jak poniżej.

```
1 t = var('t')
  x = function('x')(t)
3 equation = 5*diff(x,t,2)+diff(x,t)+8*x ==  0
  solution = desolve(equation,x,[0,0.1,0])
```

Przyjęto równanie różniczkowe zwyczajne rzędu drugiego, które zostało we wcześniejszym podrozdziale rozwiązane za pomocą Maximy. Łatwo można się przekonać, że postać rozwiązania jest dokładnie taka sama jak uzyskana poprzednio.

12.4. Podsumowanie

Niniejszy rozdział miał za zadanie zasygnalizować istnienie systemów algebry komputerowej, które mogą wspomagać pracę inżyniera. Skrótowo przedstawiono wybrane zagadnienia związane z materiałem przedstawianym w innych rozdziałach niniejszego skryptu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Przedstawione przykłady dotyczą niezwykle prostych zagadnień. Za pomocą SageMath i Maximy można rozwiązywać znacznie bardziej skomplikowane problemy. W przypadku wielu niestandardowych zagadnień wsparcie takich systemów może przyspieszyć osiągnięcie rozwiązania. Oczywiście nic nie zastąpi wykwalifikowanego matematyka, lecz oprogramowanie może przy wielu problemach ułatwić pracę.

Rozmyślnie pominięto omawianie takich pakietów jak Maple czy Mathematica, ponieważ ich dokumentacja jest obszerna i szczegółowa. Bardzo łatwo można znaleźć materiały na ich temat. Jeżeli chodzi o pakiety wolnodostępne, sytuacja jest zdecydowanie gorsza. Autorzy mają nadzieję, że ten fragment skryptu pozwoli, przynajmniej w minimalnym stopniu, zapełnić istniejącą lukę.

Jak wykazano, używanie systemów algebry komputerowej ma pewne wady. Niemniej użytkownik, który jest ich świadomy, może w dużym stopniu podnieść efektywność swej pracy.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Część III

Metody statystyczne

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 13

Metody statystyczne

13.1. Wstęp

Istniejący dualizm opisu zjawisk fizycznych wymusza na badaczach i inżynierach wykorzystywanie metod statystycznych do analizy zjawisk. Modele deterministyczne stanowią większość modeli używanych przez inżynierów. Wynika to z prostoty ich analizy oraz zgodności wyników obliczeń z obserwacjami. Modele deterministyczne oparte na wiedzy o zjawiskach fizycznych zachodzących w obiekcie dają wyniki mające prostą i oczywistą interpretację fizyczną. Niestety, nie zawsze takie modele są wystarczające. Szczególnie przy przeprowadzaniu badań doświadczalnych nie można pominąć niepewności wprowadzanej przez proces pomiaru.

Od dawna prowadzone są rozważania na temat wyższości modeli analitycznych nad numerycznymi. Modele analityczne dają możliwość przewidywania zachowania układu, pozwalają w miarę łatwo wyznaczać odpowiednie punkty pracy układu, dają informację o wrażliwości układu na zmianę parametrów wejściowych, czy pozwalają na łatwe rysowanie charakterystyk układu. Z kolei modele numeryczne dają możliwość uwzględniania zjawisk trudnych czy nawet niemożliwych do opisu analitycznego. Dzięki nim można rozwiązywać problemy, dla których nie są znane rozwiązania analityczne. Sztandarowym przykładem są zagadnienia przepływowe. Dzięki modelom analitycznym można zaprojektować maszyny przepływowe, ale optymalizacja wymaga uwzględniania zjawisk, których na obecnym etapie rozwoju nauki nie da się modelować analitycznie.

Niestety zasadniczą słabością modeli numerycznych jest ich dyskretność. Badając model numeryczny, uzyskuje się wynik (wartość wyjściową) y_1 dla zadanego zestawu parametrów wejściowych. Co więcej, najczęściej nie ma żadnej informacji na temat wrażliwości układu na zmianę parametrów w pobliżu analizowanego punktu pracy. Zmiana wartości wejściowych prowadzi do kolejnego zestawu wyników y_2, ale nie ma możliwości określenia, jak się zachowuje układ pomiędzy wartościami y_1 i y_2. Dlatego do optymalizacji trzeba używać tzw. metod bezgradientowych, czyli metod, które działają, wykorzystując informacje jedynie o wartości funkcji. Zazwyczaj metody gradientowe są znacznie szybsze, ale wymagają informacji o lokalnej wrażliwości układu na zmiany parametrów wejściowych. W ten sposób pojawia się potrzeba budowania analitycznego opisu zachowania modeli numerycznych. Taki model analityczny powinien przynajmniej lokalnie dobrze opisywać zmiany wartości wyjściowych przy zmianie wartości wejściowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Od kiedy w modelach numerycznych (komputerowych) uwzględnia się coraz więcej zjawisk fizycznych zachodzących w realnym świecie, analiza zachowania modelowanych układów staje się trudniejsza. Dzieje się tak, ponieważ pojawia się znacznie więcej zmiennych wejściowych, niezależnych i znacznie trudniej jest wyciągać ogólne wnioski na temat zachowania układu.

13.2. Pojęcia podstawowe

13.2.1. Prawdopodobieństwo

Pojęcie prawdopodobieństwa jest związane z pojęciem doświadczenia, w wyniku którego zachodzi zdarzenie. Jeżeli wynik doświadczenia nie jest jednoznacznie zdeterminowany, to mówi się o zdarzeniach losowych. Po wielokrotnym powtarzaniu doświadczenia można stwierdzić, że różne wyniki pojawiają się z określoną częstością względną. Oznacza to, na przykład, że dla określonego zdarzenia A, które może polegać na wyrzuceniu orła w doświadczeniu polegającym na rzucie monetą, przy wykonaniu N doświadczeń zdarzenie A zaszło N_A razy. Zatem względna częstość n_A zachodzenia zdarzenia A wynosi

$$n_a = \frac{N_A}{N} \tag{13.1}$$

Prawdopodobieństwo jest liczbą $P(A)$ przypisaną zdarzeniu A w następujący sposób

$$P(A) = \lim_{N\to\infty} n_a = \lim_{N\to\infty} \frac{N_A}{N} \tag{13.2}$$

Jak zatem widać, jest to idealizacja pojęcia częstości względnej zachodzenia zdarzenia A. Chociaż powyższe określenie prawdopodobieństwa nie jest ścisłą definicją matematyczną, jest wystarczające na potrzeby niniejszego skryptu. Osoby zainteresowane tematem powinny sięgnąć do podręczników z zakresu statystyki matematycznej i rachunku prawdopodobieństwa, zawierających bardziej obszerny wykład omawianych zagadnień.

Zmienna losowa rzeczywista, a o takich będzie mowa w dalszej części niniejszego podręcznika, jest funkcją, która zdarzeniom przypisuje wartości ze zbioru liczb rzeczywistych. Zmienne losowe mogą być ciągłe lub dyskretne. Przy zmiennych dyskretnych każdej wartości (odpowiadającej zdarzeniu) jest przypisywane prawdopodobieństwo zajścia danego zdarzeniu. Zmienne ciągłe opisuje się za pomocą rozkładu gęstości prawdopodobieństwa.

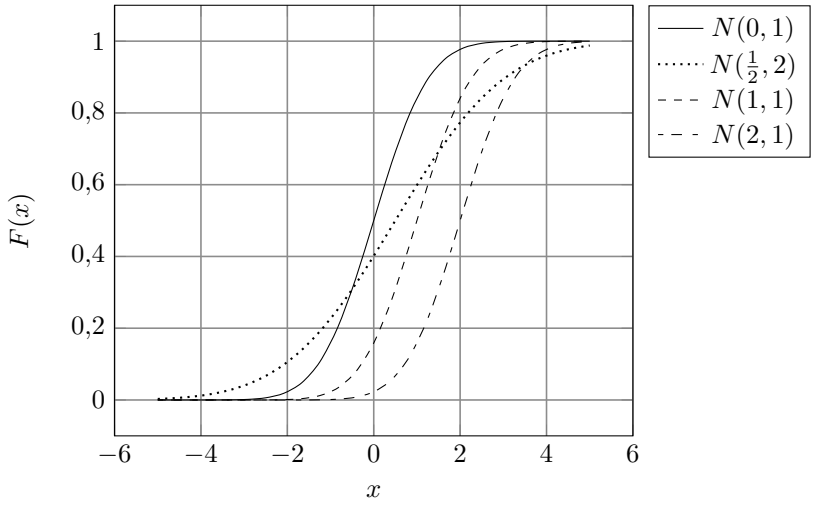
Dystrybuanta zmiennej losowej. Gdy jest dana zmienna losowa X oraz liczba rzeczywista x, to dystrybuanta zmiennej losowej jest funkcją, która określa prawdopodobieństwo zajścia zdarzenia polegającego na tym, że $X < x$.

Dystrybuanta zmiennej losowej jest funkcją monotoniczną niemalejącą, zaś jej granice wynoszą

$$\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to\infty} F(x) = 1$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Przykładowe dystrybuanty zmiennej losowej o rozkładzie normalnym zostały przedstawione na rysunku 13.1.



Rysunek 13.1. Przykładowe dystrybuanty zmiennych losowych o rozkładzie normalnym

Rozkład gęstości prawdopodobieństwa. Jeżeli dystrybuanta zmiennej losowej jest różniczkowalna, to jej pierwsza pochodna jest nazywana funkcją gęstości prawdopodobieństwa (13.3).

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = F'(x) \tag{13.3}$$

Ważną właściwością funkcji gęstości prawdopodobieństwa zmiennych losowych jest

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

Przykładowe rozkłady gęstości prawdopodobieństwa zmiennych losowych o rozkładach normalnych przedstawiono na rysunku 13.2

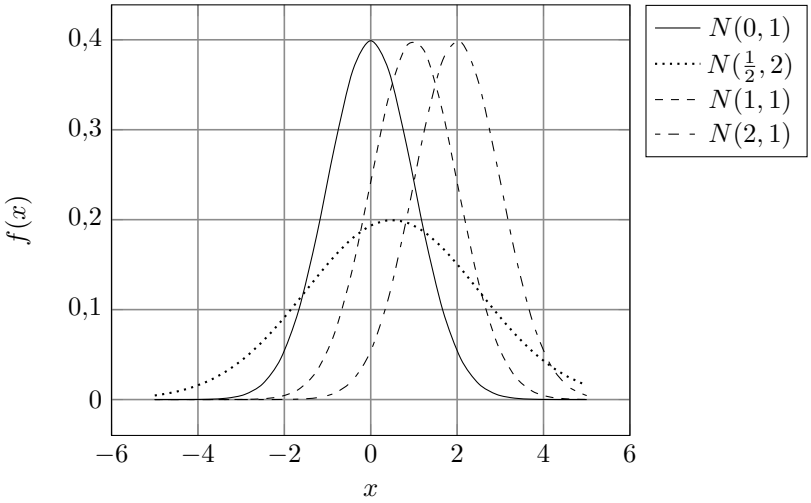
Wartość oczekiwana zmiennej losowej X oznaczana jest przez $\mathrm{E}(X)$. W przypadku, gdy X jest zmienną typu dyskretnego, wartość oczekiwana wynosi

$$\mathrm{E}(X) = \sum_i x_i p_i$$

W przypadku, gdy X jest zmienną typu ciągłego o gęstości f, wartość oczekiwana wyraża się wzorem jak poniżej i istnieje, gdy ta całka jest zbieżna.

$$\mathrm{E}(X) = \int_{\mathbb{R}} x f(x)\mathrm{d}x$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 13.2. Przykładowe rozkłady gęstości prawdopodobieństwa zmiennych losowych o rozkładzie normalnym

Wartość oczekiwana ma następujące własności:

— $X \geqslant 0 \Rightarrow \mathrm{E}(X) \geqslant 0$,
— $|\mathrm{E}(X)| \leqslant \mathrm{E}(|X|)$,
— dla $a, b \in \mathbb{R}$ zachodzi $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$,
— dla $a \in \mathbb{R}$ zachodzi $\mathrm{E}(a) = a$,
— $\mathrm{E}(X - \mathrm{E}(X)) = 0$,
— $\mathrm{E}(XY) = \mathrm{E}(X) \cdot \mathrm{E}(Y)$, gdy X i Y są niezależne.

Wartość oczekiwaną z próby nazywa się średnią arytmetyczną

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Wariancja zmiennej losowej X to liczba $\mathrm{Var}(X)$ dana wzorem

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$

W przypadku zmiennej losowej X typu dyskretnego zachodzi zależność

$$\mathrm{Var}(X) = \sum_i (x_i - \mathrm{E}(X))^2 p_i$$

Własności wariancji w skrócie zestawiono poniżej:

— $\mathrm{Var}(X) \geqslant 0$,
— $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$ dla $c \in \mathbb{R}$,

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— $\mathrm{Var}(X + c) = \mathrm{Var}(X)$,
— $\mathrm{Var}(X) = 0 \iff \exists_c P(X = c) = 1$,
— $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, gdy X i Y są niezależne.

Wariancją z próby nazywa się średnią arytmetyczną kwadratów odchyleń zmiennych losowych od wartości oczekiwanej z próby, czyli

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Pierwiastek drugiego stopnia z wariancji nazywa się odchyleniem standardowym

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

i zazwyczaj jest oznaczany przez σ.

Kowariancja jest miarą zależności liniowej między zmiennymi losowymi X i Y. Definiuje się ją jako wartość oczekiwaną z iloczynu odchyłek, czyli

$$\mathrm{cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{13.4}$$

Kowariancja wykrywa tylko możliwe zależności liniowe. Jeżeli taka zależność liniowa nie jest wykrywana, kowariancja przyjmuje wartość zero. Przy czym nie oznacza to, że zmienne są całkowicie niezależne. Mogą być zależne nieliniowo.

Współczynnik korelacji Pearsona. Podobnie jak kowariancja, jest on miarą liniowej zależności między zmiennymi losowymi. Może być traktowany jak znormalizowana kowariancja.

$$R_{XY} = \frac{\mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right)}{\sigma_X \sigma_Y} = \frac{\mathrm{cov}(XY)}{\sigma_X \sigma_Y} \tag{13.5}$$

Wartość współczynnika $R_{XY} = 1$ oznacza dodatnią korelację (rosnącą zależność liniową), zaś $R_{XY} = -1$ oznacza malejącą zależność liniową.

Należy pamiętać, że korelacja równa ± 1 nie oznacza związku przyczynowo-skutkowego, a jedynie bliżej nieokreślone przyczynowo powiązanie między wartościami zmiennych.

Hipotezę statystyczną stanowi jakiekolwiek przypuszczenie dotyczące nieznanego rozkładu zbiorowości generalnej. Przy weryfikacji hipotez zbiór hipotez dopuszczalnych, oznaczających zbiór możliwych rozkładów, dzieli się na dwa podzbiory i na podstawie próby ustala się, do którego podzbioru należy rozkład badanej zbiorowości generalnej.

Hipotezy mogą być proste i złożone. Jeżeli hipoteza dotyczy jednego konkretnego rozkładu, to jest to hipoteza prosta, w przeciwnym przypadku jest hipotezą złożoną.

Hipotezę alternatywną przyjmujemy, gdy odrzucamy hipotezę zerową.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Weryfikacja hipotez przeprowadzana jest z wykorzystaniem testów statystycznych i daje obiektywną informację o tym, czy daną hipotezę można przyjąć, czy należy ją odrzucić. Znane jest przy tym prawdopodobieństwo popełnienia błędu. Testy mogą być parametryczne, gdy znana jest postać funkcyjna rozkładu prawdopodobieństwa i poszukiwane są parametry tego rozkładu. Testy nieparametryczne służą do weryfikacji hipotez w których nie ma założeń o postaci funkcji opisującej rozkład:

— o adekwatności rozkładu,
— o tym, że dwie populacje mają różne średnie,
— o wartości średniej.

Estymacja oznacza proces zdobywania informacji na temat zbiorowości generalnej na podstawie wybranej próby. Podczas estymacji konieczna jest znajomość klasy rozkładów, do której należy nieznany rozkład badanej zbiorowości generalnej. Gdy znany jest rozkład, a poszukuje się jedynie wartości parametrów opisujących go, mówi się o estymacji parametrycznej. Gdy zaś nie jest znana postać funkcyjna rozkładu, dokonać trzeba estymacji nieparametrycznej.

Estymację można też podzielić ze względu na uzyskiwane informacje na temat szukanego rozkładu na estymację punktową, gdy na podstawie próby przyjmuje się pewne oszacowanie prawdziwej wartości parametrów rozkładu, oraz estymację przedziałową, gdy wyznacza się pewien przedział, który zawiera prawdziwą wartość poszukiwanego parametru zbiorowości generalnej.

Przy estymacji przedziałowej zakłada się prawdopodobieństwo, z którym poszukiwana wartość ma zawierać się w przedziale i na tej podstawie dobiera się szerokość przedziału. Przy estymacji przedziałowej mówi się poziomie istotności α, oznaczającym prawdopodobieństwo popełnienia błędu przy estymacji. Wielkość $1-\alpha$ nazywa się poziomem ufności, zaś przedział, w którym znajduje się poszukiwana wartość z prawdopodobieństwem równym $1-\alpha$, nazywa się przedziałem ufności. Estymatory mogą być nieobciążone, gdy jego wartość oczekiwana równa się wartości prawdziwej parametru estymowanego, lub też mogą być obciążone, gdy taki estymator wprowadza błąd systematyczny.

13.2.2. Rozkłady gęstości prawdopodobieństwa zmiennych losowych

W podrozdziale przedstawiono podstawowe informacje na temat najczęściej używanych w eksperymentach rozkładów gęstości prawdopodobieństwa ciągłych zmiennych losowych. Materiał opracowano na podstawie publikacji [17, 34].

Rozkład normalny jest jednym z ważniejszych rozkładów gęstości prawdopodobieństwa zmiennych losowych. Funkcja gęstości prawdopodobieństwa $f(x)$ wyraża się następującym wzorem

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{dla} \quad x \in \mathbb{R}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie σ oraz μ są stałe i mają interpretację, odpowiednio odchylenia standardowego oraz wartości oczekiwanej. Dystrybuanta rozkładu normalnego wyraża się wzorem

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$$

Dla rozkładu normalnego używa się oznaczenia $N(\mu, \sigma^2)$, zaś rozkład $N(0, 1)$ nazywamy standardowym rozkładem normalnym. Każdą zmienną losową o rozkładzie $N(\mu, \sigma^2)$ można poddać standaryzacji zgodnie ze wzorem

$$z = \frac{x - \mu}{\sigma}$$

Zmienna z ma tzw. standardowy rozkład normalny oznaczany $\phi(z)$, zaś jej dystrybuantę oznacza się $\Phi(z)$.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} \mathrm{d}x$$

Dystrybuanta standardowego rozkładu normalnego jest stablicowana i można ją znaleźć w każdych tablicach statystycznych. W programach obliczeniowych wśród funkcji statystycznych można znaleźć tzw. funkcję błędu, oznaczaną przez `erf(z)`

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-x^2} \mathrm{d}x$$

Za pomocą funkcji błędu można wyznaczyć wartości dystrybuanty standardowego rozkładu normalnego zgodnie z równaniem

$$\Phi(z) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right)$$

Ze zmiennych losowych o standardowym rozkładzie normalnym można uzyskać zmienne losowe o dowolnym rozkładzie normalnym. Jeżeli zmienna losowa Z ma rozkład $N(0, 1)$, to zmienna losowa $X = \sigma Z + \mu$ ma rozkład normalny $N(\mu, \sigma)$.

Rozkład χ^2 Jeżeli ciąg niezależnych zmiennych losowych $X_i \sim N(0, 1)$ ma rozkład normalny oraz

$$Y = \sum_{i=1}^{k} (X_i)^2$$

to zmienna losowa Y ma rozkład χ^2 o k stopniach swobody. Liczba stopni swobody opisuje liczbę niezależnych składników w sumie. Funkcja gęstości prawdopodobieństwa rozkładu χ^2 wyraża się zależnością

$$f(x;\, k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie $\Gamma(k/2)$ oznacza funkcję Γ Eulera, która ma postać jawną dla k całkowitego. Wartość oczekiwana rozkładu χ^2 wynosi $\mathrm{E}(x) = k$ oraz wariancja wynosi $\mathrm{Var}(x) = 2k$. Rozkład dla $k \leqslant 2$ maleje asymptotycznie, dla $k > 2$ ma maksimum w punkcie $x = k - 2$. Zgodnie z twierdzeniami granicznymi, rozkład χ^2 szybko upodabnia się do rozkładu normalnego. Szybko w tym przypadku oznacza zazwyczaj liczbę stopni swobody większą niż 30. Warto jeszcze wspomnieć, że jeżeli dwie zmienne losowe χ_1^2 o N stopniach swobody i χ_2^2 o M stopniach swobody są niezależne, to suma tych zmiennych również ma rozkład χ^2 o $N + M$ stopniach swobody.

Rozkład t-Studenta. Rozkład Studenta opracował w 1908 r. William Sealy Gosset (publikujący swe prace pod pseudonimem Student). Zmienna losowa t ma rozkład t-Studenta o k stopniach swobody, jeśli ma postać następującą

$$t = \frac{x}{\sqrt{\chi}}\sqrt{k}$$

gdzie z dwóch niezależnych zmiennych losowych, x jest zmienną losową o standardowym rozkładzie normalnym $N(0, 1)$, zaś χ jest zmienną losową o rozkładzie χ^2 o k stopniach swobody. Zmienna losowa t określona powyżej ma gęstość prawdopodobieństwa opisaną wzorem

$$f(t, k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

gdzie $\Gamma(x)$ to funkcja gamma Eulera oraz $k > 2$, zaś $-\infty < t < \infty$.

Ważność rozkładu t-Studenta wynika z tego, że dla n niezależnych zmiennych losowych o rozkładzie $N(\mu, \sigma)$ zmienna losowa

$$t_k = \frac{x_i - \bar{x}}{S_x}\sqrt{n-1}$$

ma rozkład t-Studenta o $k = n - 1$ stopniach swobody. Wartości $\bar{x}$ oraz S_x należy obliczyć zgodnie z wzorami $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oraz $S_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Rozkład t-Studenta jest symetryczny względem osi $t = 0$, więc wartość oczekiwana jest równa zeru, zaś wariancja jest równa

$$\mathrm{Var}(t) = \sigma^2(t) = \frac{k}{k-2}$$

Dla liczby stopni swobody $k > 30$ rozkład t-Studenta jest praktycznie zgodny ze standardowym rozkładem normalnym $N(0, 1)$, natomiast odchylenie standardowe jest większe niż w rozkładzie normalnym, co wizualnie oznacza, że jego funkcja gęstości wolniej zmierza do zera.

Rozkład F-Snedecora. Zmienna losowa z ma rozkład F-Snedecora lub w skrócie rozkład F, jeżeli

$$z = \frac{X}{Y}\frac{k_2}{k_1}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

gdzie obie zmienne losowe mają rozkład χ^2 oraz X ma k_1 stopni swobody, zaś Y ma k_2 stopni swobody. Funkcja gęstości rozkładu prawdopodobieństwa wyraża się wzorem

$$f(z) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_2}{k_1}\right)^{\frac{k_2}{2}} z^{\left(\frac{k_2-2}{2}\right)} \left(1+\frac{k_2}{k_1}z\right)^{-\frac{k_1+k_2}{2}}$$

Chociaż w różnych źródłach można znaleźć różne postacie zależności opisujących rozkład F, dają one te same kształty.

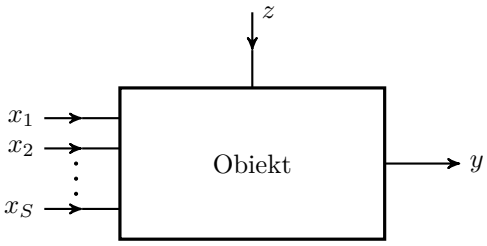
Poniżej zebrano polecenia z pakietów MATLAB i GNU Octave służące do generowanie opisanych rozkładów gęstości prawdopodobieństwa (*probability density function*).

Rozkład	Polecenie
Rozkład normalny	`normpdf`, `std_normal`
Rozkład χ^2	`chi2pdf`
Rozkład t-Studenta	`tpdf`
Rozkład F	`fpdf`

Zazwyczaj podaje się parametry rozkładu i zbiór wartości zmiennej niezależnej, dla których należy wyznaczyć wartości rozkładu gęstości prawdopodobieństwa. Szczegóły można znaleźć w pomocy pakietów. Aby korzystać z wyżej wymienionych poleceń w GNU Octave, należy najpierw wczytać pakiet statystyczny poleceniem `pkg load statistics`.

13.2.3. Obiekt i jego model

Podczas przeprowadzania badań eksperymentalnych obecny jest pewien obiekt badań, który poddawany jest kolejnym doświadczeniom [17, 34–36] (rys. 13.3).



Rysunek 13.3. Opisywany obiekt nieliniowy

W poddany eksperymentowi obiekcie jest S wejść $x_1, x_2, \ldots, x_S$, określanych jako czynniki lub zmienne objaśniające oraz N wyjść $y_1, y_2, \ldots, y_N$ zwanych odpowiedziami obiektu. Ponadto badany obiekt poddawany jest nieznanym oraz niemierzalnym

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

zakłóceniom $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$. Celem przeprowadzenia eksperymentu jest doprowadzenie do stworzenia modelu matematycznego analizowanego obiektu. Dlatego podczas badań wartości zmiennych objaśniających x_i są modyfikowane w celu obserwacji odpowiedzi obiektu y. Zakłócenia ε są przypadkowe i podlegają nieprzewidywalnym zmianom. Powtarzając zatem eksperyment przy tych samych określonych wartościach wejść, można otrzymać różne wartości wyjść. Problem ten nie dotyczy jednak eksperymentów cyfrowych, co stanowi ich zaletę. Jako że związek pomiędzy odpowiedziami obiektu a czynnikami jest zwykle nieznany, to badany obiekt jest opisywany za pomocą nieznanej, ogólnie nieliniowej funkcji, której argumentami są określone czynniki x_i, a także niemierzalne zakłócenia ε_i

$$y = f(x_1, x_2, \ldots, x_S, \varepsilon_1, \varepsilon_2, ..., \varepsilon_N) \tag{13.6}$$

Jak już wspomniano, ponieważ rzeczywista postać powyższej funkcji nie jest znana, obiera się model matematyczny będący funkcją aproksymującą odpowiedzi obiektu w badanym obszarze. Należy wspomnieć, że odpowiedni model matematyczny jest wyznaczany na podstawie obserwacji wyjść y układu.

Wszystkie obserwacje wyjść y obiektu są obarczone pewnym błędem ε. W praktyce, wyróżnia się dwa rodzaje błędów:

— systematyczny, który może być związany z niewłaściwym skalibrowaniem aparatury, błędami popełnionymi w trakcie przygotowań itp; jego występowanie jest związane z błędami popełnianymi przez osobę wykonującą dany eksperyment;
— losowy, który jest niezależny od eksperymentatora i występuje zawsze.

W przypadku błędów losowych zakłada się, że wszelkie zakłócenia ε są niezależnymi zmiennymi losowymi o rozkładzie normalnym z zerową wartością oczekiwaną i jednakowych dla wszystkich zakłóceń wariancjach σ^2

$$\varepsilon_i = N(0, \sigma^2) \tag{13.7}$$

Ponadto, zakłada się, że zakłócenia te oddziałują addytywnie na obiekt badań

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \ldots, N \tag{13.8}$$

gdzie μ_i są prawdziwymi, nieznanymi wartościami wyjść modelu.

13.2.4. Optymalizacja

Z punktu widzenia matematyka optymalizacja jest procesem poszukiwania minimum (ogólniej ekstremum) pewnej funkcji z uwzględnieniem pewnych ograniczeń, zaś w praktyce inżynierskiej jest poszukiwaniem najlepszego, ze względu na przyjęte kryterium, rozwiązania spośród dopuszczalnych. Te podejścia, chociaż mogą na pierwszy rzut oka być odległe, są tożsame. W rzeczywistości inżynier przyjmuje pewną funkcję celu, która każdemu rozwiązaniu przypisuje jakąś wartość, zaś optymalizacja sprowadza się do poszukiwania minimum (ekstremum) tej funkcji. Zatem proces optymalizacji można sprowadzić do poszukiwania wartości minimalnej lub maksymalnej wyjścia obiektu w ogólnym przypadku nieliniowego (patrz rysunek 13.3) [17, 34–36]. Czyli

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

poszukuje się takiego punktu x_{opt}, że w jego otoczeniu wartości funkcji $f(x) = y(x)$ są większe od $f(x_{opt})$ lub równe, co można zapisać

$$f(x_{opt}) \leqslant f(x), \text{ dla } |x_{opt} - x| \leqslant \varepsilon, \ (\varepsilon > 0)$$

Powyższa definicja może się odnosić do maksymalizacji funkcji, należy jedynie dokonać zmiany znaku funkcji $f(x) = -f(x)$. Chociaż definicja wydaje się mocno abstrakcyjna, to należy pamiętać, że od umiejętności właściwego wyboru optymalizowanej funkcji często zależy pozytywny wynik procesu.

13.2.5. Eksperyment i jego plan

W niniejszym skrypcie przez eksperyment rozumie się serię doświadczeń, umożliwiającą stworzenie opisu matematycznego analizowanego obiektu [34]. Budowa modelu matematycznego służy poznaniu obiektu lub poprawie jego działania. Czyli celem eksperymentu jest identyfikacja lub optymalizacja rozważanego obiektu.

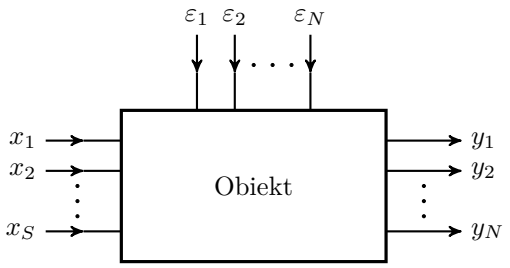
Eksperyment może być bierny, gdy obserwacje są zbierane podczas pracy obiektu, a eksperymentator nie ma możliwości modyfikowania wartości wejściowych. Często w eksperymencie biernym uzyskuje się wyniki jakościowe. Można też przeprowadzać eksperyment czynny, którego dotyczy ta część książki, polega on na badaniu zachowania obiektu pod wpływem założonych zmian wartości czynników wejściowych.

Ponieważ w większości przypadków przeprowadzenie eksperymentu nie jest tanie, z reguły dąży się do wykonania możliwie krótkich serii doświadczeń. Należy też pamiętać, że zarówno jakość identyfikacji, jak i optymalizacji zależą od doboru parametrów wejściowych do doświadczeń. Wobec tego należy odpowiednio zaplanować wartości wymuszeń. Podstawą metody planowania eksperymentu są dwie metody identyfikacji, znane pod nazwami analizy czynnikowej oraz analizy regresji.

Eksperymenty mogą dotyczyć układów o jednym wejściu i jednym wyjściu, lub wielu wejściach i wyjściach. Nic nie tracąc na ogólności rozważań, zakłada się, że układ o wielu wyjściach da się przedstawić w postaci szeregu układów o jednym wyjściu. W ten sposób rozważane obiekty należą do klasy układów przedstawionych na rysunku 13.3, na których wyjścia oddziałują, oprócz zadanych wartości wejściowych, nieznane zakłócenia. Eksperyment jest rozumiany jako seria doświadczeń na obiekcie rzeczywistym lub doświadczeń cyfrowych. Schematycznie eksperyment może zatem zostać przedstawiony w postaci układu o wielu wejściach i wielu wyjściach (patrz rysunek 13.4).

Eksperyment ma dostarczyć informacji pozwalających na zbudowanie modelu matematycznego badanego obiektu. Ten model może wykorzystywać wiedzę o zasadzie działania obiektu i wynikać z podstawowych praw fizyki. Może też być arbitralnie przyjętą funkcją, opisującą zależność między wartościami wejść a wartością uzyskaną na wyjściu obiektu. Jeżeli przy budowie modelu była brana pod uwagę struktura fizyczna modelu i związane z nią prawa fizyki, eksperyment może dawać informacje mające interpretację fizyczną. W przeciwnym przypadku uzyskane zależności mogą mieć tylko charakter korelacyjny, nic niemówiący o zależnościach przyczynowo-skutkowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 13.4. Schematyczna reprezentacja eksperymentu

Eksperyment ma pozwolić na identyfikację współczynników modelu. Czyli ma pozwolić na wyznaczenie ich wartości liczbowych. Aby osiągnąć tak postawiony cel, należy zrealizować kilka etapów, które są określane ogólną nazwą planowania eksperymentu.

Pierwszy etap planowania eksperymentu polega na wyborze postaci modelu matematycznego obiektu. Musi on uwzględniać wszystkie istotne z punktu widzenia eksperymentatora zmienne wpływające na wartość wyjściową i może, ale nie musi, wykorzystywać wiedzę o fizycznej strukturze obiektu. W następnym podrozdziale znajduje się tabela, która ułatwia wybór zależności, którą można opisać wyniki eksperymentu.

Dobry eksperyment powinien spełniać pewne warunki, z których najistotniejsze to:

— brak błędu systematycznego,
— mała wartość błędu pomiarowego,
— szeroki zakres zastosowania wniosków wynikających z eksperymentu,
— prosty plan doświadczenia,
— ścisłe określenie niepewności wniosków z niego wynikających.

Szerszy opis tych i innych wymagań dotyczących eksperymentów można znaleźć w pozycji [17]. Oprócz wymienionych cech eksperyment powinien spełniać pewne kryteria jakościowe, które zebrano poniżej. Dotyczą one głównie samego planu eksperymentu i pozwalają wybrać taki, który przyczyni się do powodzenia badań. Choć wymienione poniżej kryteria są istotne podczas planowania eksperymentu, należy pamiętać, że nie istnieje plan, który jednocześnie spełniałby wszystkie stawiane wymagania.

Zatem plan eksperymentu powinien:

— być ortogonalny,
— mieć korzystne cechy numeryczne,
— mieć symetrię obrotową, innymi słowy być rotatabilny,
— być w jakimś sensie optymalny,
— minimalizować czas i koszty eksperymentu,
— w miarę możliwości uwzględniać korelację zakłóceń,
— wykorzystywać informację a priori o parametrach,
— być sekwencyjny, gdy obiekt wykazuje silnie nieliniowe zachowanie.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Pod pojęciem planu ortogonalnego rozumie się taki plan, w którym kolumny macierzy wejść są względem siebie ortogonalne, czyli ich iloczyn skalarny jest równy zero. Pominięcie w modelu pewnych członów nie powoduje konieczności przeliczania oszacowań pozostałych parametrów, jeśli pomiary wykonywane były zgodnie z planem ortogonalnym dla tego modelu. Powyższe stwierdzenie sprawia, że ze statystycznego punktu widzenia plany te mają wiele zalet. Można też powiedzieć, że macierz informacyjna $\mathbb{X}^T\mathbb{X}$ jest diagonalna.

Jeżeli chodzi o walory numeryczne planu eksperymentu, to należy oczekiwać, że jego użycie będzie prowadziło do redukcji błędów numerycznych powstających przy obliczaniu oszacowań parametrów modelu, na przykład przy zastosowaniu metody najmniejszych kwadratów.

Symetria obrotowa planu decyduje o stałości wariancji oszacowania wyjścia modelu w stałej odległości od centrum planu. Przez centrum planu rozumie się taki punkt w przestrzeni wejść, w którego otoczeniu tworzony jest model matematyczny procesu. Celem symetrii obrotowej jest zapewnienie jednakowej dokładności oszacowania wartości wyjścia modelu w każdym kierunku. Plany takie nazywa się rotatabilnymi.

Plan eksperymentu może być w jakimś sensie optymalny. Kryterium to oznacza, że przyjęty został pewny wskaźnik mierzący jakość różnych planów. Przykładowo, może to być dokładność oszacowania parametrów modelu. Mówiąc inaczej, plan optymalny to taki, który zapewnia największą możliwą do osiągnięcia w danych warunkach wartość przyjętego wskaźnika. Najpopularniejsze są plany optymalne typu D, E, A oraz G [34], przy czym:

— plan D-optymalny minimalizuje objętość elipsoidy koncentracji rozkładu ocen współczynników regresji i sprowadza się do minimalizacji wyznacznika macierzy kowariancyjnej $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$;

— plan E-optymalny to taki, w którym długość największej osi elipsoidy koncentracji jest minimalna, czyli należy maksymalizować najmniejszą wartość własną unormowanej macierzy informacyjnej $M = \frac{1}{N}\mathbb{X}^T\mathbb{X}$;

— plan A-optymalny minimalizuje średnią wariancję współczynników regresji i sprowadza się do minimalizacji śladu macierzy kowariancyjnej $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$;

— plan G-optymalny minimalizuje największą wariancję prognozowanej wartości funkcji regresji $\hat{y}(\mathbb{x})$.

Pokreślić należy, że plany D-optymalne są jednocześnie G-optymalne.

Minimalizacja czasu i kosztów eksperymentu nie jest wymaganiem łatwym do spełnienia. Przeważnie kryterium to jest spełniane jedynie pośrednio, przez dążenie do minimalizacji liczby przeprowadzanych doświadczeń.

W klasycznej teorii planowania eksperymentu zakłada się brak korelacji zakłóceń między kolejnymi pomiarami nawet wtedy, gdy dotyczą tych samych stanów wejść obiektu, chociaż w obiektach rzeczywistych taka korelacja czasem występuje.

Plany wykorzystują informację *a priori* o parametrach współczynników regresji. Jest to tak zwane podejście bayesowskie. Jego zastosowanie wymaga założenia, że nieznane wartości parametrów modelu są zmienną losową o znanym rozkładzie prawdopodobieństwa niekoniecznie normalnym.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Planowanie sekwencyjne stosuje się, gdy obiekt zachowuje się silnie nieliniowo i konieczne jest zmienianie zakresów wartości przyjmowanych przez zmienne wejściowe.

13.3. Proste modele matematyczne obiektów

W bardzo wielu przypadkach da się opisać badane obiekty stosunkowo prostymi zależnościami funkcyjnymi. Często te zależności mają uzasadnienie teoretyczne i oczywistą interpretację fizyczną. Najprostszą taką zależnością jest zależność liniowa. Jej intuicyjna zrozumiałość oraz łatwość wyznaczania współczynników powoduje, że jest też najwygodniejszą postacią zależności funkcyjnej między zmiennymi. Dlatego przy badaniach eksperymentalnych dobrze jest dopasowywać uzyskiwane wyniki do takich prostych modeli. Na podstawie tej wiedzy teoretycznej zazwyczaj można proponować opis matematyczny obiektu. Często zależności teoretyczne da się sprowadzić do pewnej klasy funkcji. Wybrane równania, które można stosować do opisu danych doświadczalnych zebrano w tabeli 13.1. Te zależności zaczerpnięto z pozycji bibliograficznych [17, 27]. Wprowadzone zmienne u oraz v oznaczają zmienne x oraz y po transformacji, która prowadzi do wyprostowania zależności funkcyjnej.

Tabela 13.1. Przekształcenia do prostowania wybranych zależności opisujących dane eksperymentalne

Lp.	Równanie	Sposób wyprostowania	Uwagi
1	$y = \frac{a}{x} + b$	$u = \frac{1}{x}$	
2	$y = a + b \ln x$	$u = \ln x$	
3	$y = ax^b$	$u = \ln x,\ v = \ln y$	
4	$y = ae^{bx}$	$v = \ln y$	
5	$y = ae^{b/x}$	$v = \ln y\ u = \frac{1}{x}$	
6	$y = \frac{1}{a+bx}$	$v = \frac{1}{y}$	
7	$y = \frac{x}{a+bx}$	$v = \frac{x}{y}$	
8	$y = c + ax^b$	$\begin{cases} \text{znane } b: & u = x^b \\ \text{znane } c: & u = \ln x, \\ & v = \ln(y-c) \end{cases}$	$x_3 = \sqrt{x_1 x_2}$, $c = \frac{y_1 y_2 - y_3^2}{y_1 + y_2 - 2y_3}$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 13.1. cd.

9	$y = c + ae^{bx}$	wyznaczyć c, $v = \ln(y - c)$	$x_3 = y_3 \frac{x_1+x_2}{2}$, $c = \frac{y_1 y_2 - y_3^2}{y_1+y_2-2y_3}$
10	$y = a + bx + cx^2$	$v = \frac{y-y_1}{x-x_1}$	równanie prostej: $v = (b + cx_1) + cx$
11	$y = \frac{a+bx}{c+dx}$	$v = \frac{x-x_1}{y-y_1}$	równ. $v = A + Bx$, A i B wyznaczyć z $y = y_1 + \frac{x-x_1}{A+Bx}$
12	$y^2 = a + bx + cx^2$	$z = y^2$	jak pozycja 10
13	$y = ae^{bx+cx^2}$	$z = \ln y$	jak pozycja 10
14	$\ln y = \ln a + bx + cx^2$	$z = \ln y$	jak pozycja 10
15	$y = \frac{1}{a+bx+cx^2}$	$z = \frac{1}{y}$	jak pozycja 10
16	$y = \frac{x}{a+bx+cx^2}$	$z = \frac{x}{y}$	jak pozycja 10
17	$y = a + \frac{b}{x} + \frac{c}{x^2}$	$z = \frac{1}{x}$	jak pozycja 10
18	$y = a + b \ln x + c \ln^2 x$	$z = \ln x$	jak pozycja 10
19	$y = ae^{bx^c}$	$z = \ln y$	jak pozycja 8
20	$y = ae^{be^{cx}}$	$z = \ln y$	jak pozycja 9

Użyte w tabeli 13.1, we wzorach krzywych, oznaczenia a, b, c d są stałymi. W tabeli czasem występują punkty (x_1, y_1) oraz (x_2, y_2), które są dowolnie wybranymi punktami na analizowanej krzywej. Tam gdzie wymagane są dwa punkty, by wyznaczyć trzeci x_3, dobrze jest wybierać punkty x_1 i x_2 możliwie daleko od siebie.

13.4. Wybór zmiennych wejściowych

Przy eksperymentalnym badaniu obiektów wybór czynników często jest oczywisty. Zmienne wejściowe, które mają wpływ na wartość wyjścia obiektu, są znane. Ale zazwyczaj nie jest to wiedza pełna. Gdyby wiedza o obiekcie była pełna, nie byłoby potrzeby przeprowadzania badań. Z tego powodu zawsze należy przemyśleć wybór zmiennych wejściowych, które będą używane w eksperymencie. Podstawowym wyma-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

ganiem jest to, by na te zmienne można było wpływać. Innym wymaganiem jest to, by miały istotny wpływ na wartość zmiennej wyjściowej.

Co to znaczy istotny wpływ? Eksperymentator ocenia, czy wpływ danej zmiennej jest istotny, czy też nie. Pomocne mogą być poniższe uwagi dotyczącej bilansu losowego. Szersze omówienie tej metody zawierają np. publikacje [17, 27].

Aby wskazać czynniki istotne w danym eksperymencie, konieczne jest przeprowadzenie eksperymentu wstępnego. W eksperymencie wstępnym przyjmuje się zmiany parametrów wejściowych na dwóch poziomach, najlepiej skrajnych. Następnie należy postępować zgodnie z poniższym tokiem:

— wybrać losowo z planu pełnego dla zadanej liczby czynników wiersze lub wykorzystać plan Placketta–Burmana (patrz podrozdział 14.5);
— na osi X wykresu zaznaczyć położenia czynników, oddzielnie dla poziomu wysokiego danego czynnika i dla poziomu niskiego;
— wybrać te wiersze, w których czynnik x_i znajduje się na poziomie wysokim i zaznaczyć uzyskane wartości nad pozycją dotyczącą danego czynnika i poziomu;
— wyznaczyć mediany (wartości środkowe) oddzielnie dla każdego poziomu czynnika;
— czynnik, w których różnica median między poziomami będzie największa, jest czynnikiem najistotniejszym.

Aby wyznaczyć następny w kolejności czynnik o największym wpływie, należy od wyników uzyskanych na poziomie górnym czynników odjąć różnicę median wyznaczonego czynnika o największym wpływie. W podobny sposób można określać wpływ poszczególnych interakcji.

Przykład wyboru czynnika o największym wpływie

Sformułowanie problemu. Dla obiektu z 3 wejściami wyznaczyć czynnik, który ma największy wpływ na wartość wyjścia.

Rozwiązanie. Plan eksperymentu zostanie zbudowany przez losowe wybranie 7 wierszy z pełnego dwupoziomowego planu eksperymentu dla 3 czynników.

W tabeli 13.2 przedstawiony jest plan eksperymentu oraz wyniki uzyskane w każdym z doświadczeń. Znakiem + zaznaczono poziom wysoki zmiennej, znakiem − poziom niski.

Zaznaczenie wartości wyjściowych dla każdego poziomu zmiennej i odpowiednie mediany zostały przedstawione na rysunku 13.5.

Jak widać na rysunku 13.5, bez uwzględnienia interakcji między czynnikami, największy wpływ na wartość wyjściową ma zmienna x_1. Ten wniosek wynika, z największej różnicy median dla zmiennej na niskim i wysokim poziomie.

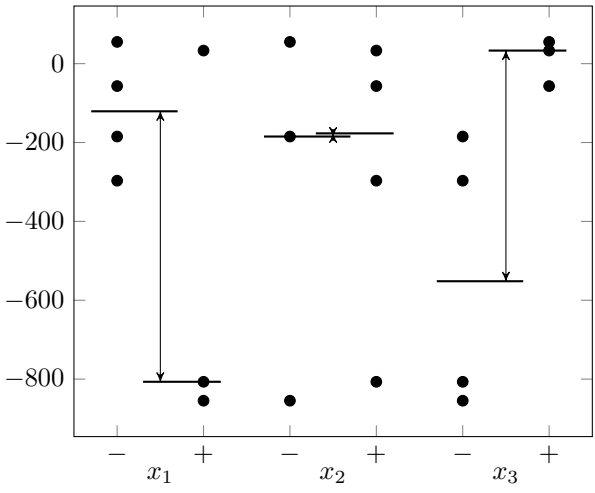
13.5. Analiza regresji

Analiza regresji pozwala na budowanie modeli matematycznych układów, na temat których wiadomo, że mają S wejść x_i, nazywanych też czynnikami, i jedno wyjście y.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 13.2. Plan i macierz eksperymentu

Lp.	x_1	x_2	x_3	y
1	−	−	−	−184,8
2	−	+	−	−296,8
3	+	−	−	−854,8
4	+	+	+	33,2
5	−	+	+	−56,8
6	−	−	+	55,2
7	+	+	−	−806,8



Rysunek 13.5. Metoda bilansu losowego

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wartość wyjścia zależy od stanów wejść oraz nieznanego i niemierzalnego zakłócenia z, które może być różne dla poszczególnych wejść. Oznacza to, że przy tych samych wartościach wejściowych wartość wyjścia będzie się zmieniać w sposób trudny do przewidzenia. Poszukiwany nieznany nieliniowy opis matematyczny może zostać zapisany w postaci

$$y = f_z(x_1, x_2, \ldots, x_S, z) \quad \text{lub} \quad y = f(x_1, x_2, \ldots, x_S) + \varepsilon \tag{13.9}$$

gdzie ε jest efektem nieznanych zakłóceń z. Ten prosty zapis matematyczny z równania (13.9) może nastręczać pewnych trudności, dlatego że wybrany do eksperymentu model matematyczny powinien spełniać następujące warunki:

— być możliwie prosty;
— charakteryzować się dużą dokładnością w punktach, w których przeprowadza się pomiary;
— zapewniać możliwość aproksymacji wartości wyjścia w punktach, w których pomiary nie zostały wykonane, ale należą do badanego obszaru;
— opis matematyczny powinien odzwierciedlać przynajmniej cechy jakościowe badanego procesu.

Z tego względu do opisu układów używa się różnych rodzin funkcji. Często o wyborze postaci modelu decyduje jego przeznaczenie oraz wielkość obszaru, w którym ma on pozwalać na dokładny opis obiektu rzeczywistego. Czasem istotna jest też łatwość wyznaczania ekstremów.

Ponieważ zakłócenie wpływające na wartość wyjścia obiektu ma charakter losowy, zatem zależność (13.9) nie jest zwykłą zależnością funkcyjną, a zależnością stochastyczną. W odróżnieniu od funkcji, w zależnościach stochastycznych nie ma jednoznaczności między wartością wejściową x a wartością wyjściową y. Co oznacza, że dla tych samych wartości x_i na wejściach do układu mogą się pojawiać różne wartości y na wyjściu z układu. Zatem istotą zależności stochastycznej jest jednoznaczna zależność rozkładu prawdopodobieństwa zmiennej losowej pojawiającej się na wyjściu od wartości zmiennej losowej podawanej na wejście układu. Zatem przy wyznaczaniu opisu matematycznego najistotniejsze będzie określenie wartości oczekiwanej zmiennej losowej Y przy warunku $X = x$ opisującym wartości podawane na wejścia układu

$$\mathrm{E}(y) = \mathrm{E}(f(x_1, x_2, \ldots, x_S)) \tag{13.10}$$

Porównanie zależności (13.9) oraz (13.10) pozwala ujawnić założenie, które zostało poczynione, a dotyczy wartości oczekiwanej nieznanych zakłóceń, która ma być równa zeru $\mathrm{E}(\varepsilon) = 0$. Oczywiście analizę regresji można stosować nie tylko przy zależnościach stochastycznych. Zależności funkcyjne też mogą być przybliżane w ten sposób. W takiej sytuacji mówi się po prostu o aproksymacji punktowej wielomianowej znanej z rozdziału 8.

Zagadnienie regresji można postawić w następujący sposób: dla S znanych wartości wejść x_1, x_2, …, x_S oraz odpowiadających im wartości wyjścia y obiektu przedstawionego na rysunku 13.3 i opisanego równaniem (13.11), wyznaczyć $K+1$ współczyn-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

ników arbitralnie przyjętego zastępczego modelu matematycznego w postaci opisanej równaniem (13.12).

$$y = f(x_1, x_2, \ldots, x_S; \beta_0, \beta_1, \ldots, \beta_K) + \varepsilon \tag{13.11}$$

$$\hat{y} = f(x_1, x_2, \ldots, x_S; b_0, b_1, \ldots, b_K) \tag{13.12}$$

Wartości nieznanych współczynników β_i równania (13.11) opisującego obiekt zostaną wyznaczone tak, by wartości oczekiwane $\mathrm{E}(y)$ oraz $\mathrm{E}(\hat{y})$ były równe, czyli $\mathrm{E}(y) = \mathrm{E}(\hat{y})$. W celu ich rozróżnienia ich wyestymowane wartości oznaczane są przez b_i. Aby ułatwić czytanie, rozróżnienie to nie będzie przestrzegane zbyt rygorystycznie w dalszej części rozważań, a będzie przywoływane, jeżeli będzie to niezbędne dla jasności wywodów.

Aby wyznaczyć $K + 1$ współczynników, konieczne jest przeprowadzenie przynajmniej $N \geqslant K+1$ pomiarów wartości wyjściowej przy różnych wartościach wejściowych. Co więcej, aby można było wyznaczyć współczynniki modelu (13.12), konieczne jest określenie miary pozwalającej określić „odległość" modelu od obiektu (13.9). Czyli należy przyjąć jakaś miarę S_R w postaci (13.13).

$$S_R = S_R\left(\{y_1, y_2, \ldots, y_N\}, \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}\right) \tag{13.13}$$

Z wielu względów bardzo dogodnym wskaźnikiem jakości dopasowania modelu do obiektu jest norma Euklidesa, która w tym przypadku wyraża się zależnością

$$S_R = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - f(x_1, x_2, \ldots, x_S; b_0, b_1, \ldots, b_K))^2 \tag{13.14}$$

Miarę wprowadził w 1806 r. Adrien-Marie Legendre i nieco później, niezależnie od niego Carl Friedrich Gauss (1809 r.). Parametry modelu (13.12) wyznacza się, minimalizując miarę (13.14), a otrzymany model nazywa się funkcją regresji. Po wprowadzeniu oznaczeń $\mathbb{x} = [x_1, x_2, \ldots, x_S]$ oraz $\mathbb{b} = [b_0, b_1, \ldots, b_K]^T$ można zapisać ogólną poszukiwaną postać funkcji regresji, liniowej względem nieznanych parametrów b_i

$$\hat{y} = \hat{y}(\mathbb{x}, \mathbb{b}) = b_0 + b_1 f_1(\mathbb{x}) + b_2 f_2(\mathbb{x}) + \ldots + b_K f_K(\mathbb{x}) \tag{13.15}$$

przy czym funkcje $f_i(\mathbb{x})$ są znane, liniowo niezależne i mogą być nieliniowe. Na koniec należy zaznaczyć, że uzyskane na drodze regresji modele matematyczne są ograniczone do zakresu badanego obszaru. Oznacza to, że nie wolno w żadnym przypadku wyciągać jakichkolwiek wniosków na temat zachowania układu dla wartości wejściowych spoza zakresu, w jakim mieściły się zmienne $\mathbb{x}$.

13.5.1. Wyznaczanie parametrów funkcji regresji

Równanie opisujące obiekt z rysunku 13.3 ma postać następującą

$$y = y(\mathbb{x}, \beta) + \varepsilon = \beta_0 + \sum_{i=1}^{K} \beta_i f_i(\mathbb{x}) + \varepsilon \tag{13.16}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Przyjęty model badanego obiektu ma postać (13.15). Czyli jest liniowy względem nieznanych parametrów b_i. W ogólnym przypadku funkcja regresji może być nieliniowa i wielowymiarowa. Nieliniowość jest ukryta w funkcjach f_i. Zadawane wartości parametrów wejściowych można wygodnie zapisać w postaci macierzy wejść $\mathbb{X}$, której wiersze opisują wymuszenia w kolejnych doświadczeniach w ramach eksperymentu.

$$\mathbb{X} = \begin{bmatrix} f_0(\mathbb{x}_1) & f_1(\mathbb{x}_1) & f_2(\mathbb{x}_1) & \dots & f_K(\mathbb{x}_1) \\ f_0(\mathbb{x}_2) & f_1(\mathbb{x}_2) & f_2(\mathbb{x}_2) & \dots & f_K(\mathbb{x}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_0(\mathbb{x}_N) & f_1(\mathbb{x}_N) & f_2(\mathbb{x}_N) & \dots & f_K(\mathbb{x}_N) \end{bmatrix} \tag{13.17}$$

Dla spójności zapisu przyjęto, że $f_0(\mathbb{x}) \equiv 1$. Przy czym $\mathbb{x}_i = [x_{i1}, x_{i2}, \dots, x_{iS}]$. Równanie (13.17) dotyczy nieliniowej funkcji regresji. Gdy model układu jest liniowy, macierz wejść ma postać (13.18), gdzie wiersze zawierają wartości x_{ij} tworzące wektory $\mathbb{x}_j$ opisujące wymuszenia w j-tych doświadczeniach czy chwilach czasowych

$$\mathbb{X} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \dots & x_{1S} \\ x_{20} & x_{21} & x_{22} & \dots & x_{2S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N0} & x_{N1} & x_{N2} & \dots & x_{NS} \end{bmatrix} \tag{13.18}$$

przy czym elementy $x_{i0} \equiv 1$ dla $i = 1, 2, \dots, N$ wprowadzono dla ujednolicenia zapisu.

Zaobserwowane wartości wyjścia z obiektu i z modelu zostaną zapisane w postaci wektorów wyjść obiektu oraz wartości funkcji regresji

$$\mathbb{y} = [y_1, y_2, \dots, y_N]^T \tag{13.19}$$

$$\hat{\mathbb{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N]^T \tag{13.20}$$

Jak już wspomniano, nieznane współczynniki poszukiwanej funkcji regresji oraz ich estymatory można zapisać wektorowo jako

$$\beta = [\beta_0, \beta_1, \dots, \beta_K]^T, \quad \mathbb{b} = [b_0, b_1, \dots, b_K]^T \tag{13.21}$$

Zatem wartości funkcji regresji oraz wartości wyjścia z obiektu dla poszczególnych doświadczeń czy chwil czasowych można zapisać macierzowo jako

$$\mathbb{y} = \mathbb{X}\beta + \varepsilon, \quad \hat{\mathbb{y}} = \mathbb{X}\mathbb{b} \tag{13.22}$$

Minimalizacja sumy kwadratów błędów będzie się sprowadzać do znalezienia minimum wyrażenia

$$S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\mathbb{y} - \mathbb{X}\mathbb{b})^T (\mathbb{y} - \mathbb{X}\mathbb{b}) = \mathbb{y}^T\mathbb{y} - 2\mathbb{b}^T\mathbb{X}^T\mathbb{y} + \mathbb{b}^T\mathbb{X}^T\mathbb{X}\mathbb{b} \tag{13.23}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Wyznaczenie pochodnej S_R i przyrównanie do zera prowadzi do układu równań normalnych Gaussa

$$\mathbb{X}^T\mathbb{X}\mathbb{b} = \mathbb{X}^T\mathbb{y} \tag{13.24}$$

zaś symboliczne rozwiązanie tego układu można zapisać jako

$$\mathbb{b} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{y} \tag{13.25}$$

Oczywiście jest to zapis czysto formalny. Można go wykorzystać do prowadzenia obliczeń, jednak zdecydowanie wydajniej jest rozwiązywać układ równań normalnych (13.24), co pozwoli też ograniczyć błędy zaokrągleń. Warto pamiętać, że estymatory b_i wyznaczane metodą najmniejszej sumy kwadratów są estymatorami współczynników β_i o najmniejszej wariancji. Fakt ten wyjaśnia popularność tej metody wyznaczania parametrów funkcji regresji.

13.5.2. Jakość funkcji regresji

W poprzednim podrozdziale wyprowadzono zależność pozwalającą na wyznaczenie współczynników funkcji regresji, przy założeniu, że dobrze opisuje ona zachowanie analizowanego układu. Aby uściślić rozważania, należy stwierdzić, że zachodzi zasadnicza różnica między równaniem opisującym zachowanie obiektu a jego modelem matematycznym w postaci funkcji regresji. Równanie obiektu ma następującą postać

$$\mathbb{y} = \mathbb{X}\beta + \varepsilon \tag{13.26}$$

gdzie ε jest wektorem nieznanych zakłóceń stochastycznie niezależnych, o rozkładzie normalnym o wartości oczekiwanej $\mathrm{E}(\varepsilon) = 0$ i macierzy kowariancji równej

$$\mathrm{cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \mathbb{I}\sigma^2$$

Wielkość wyjściowa będzie zatem zmienną losową o rozkładzie normalnym i wartości oczekiwanej równej $\mathrm{E}(\mathbb{y}) = \mathbb{X}\beta$. Natomiast uzyskany model matematyczny zapisać można zależnością

$$\hat{\mathbb{y}} = \mathbb{X}\mathbb{b} \tag{13.27}$$

Współczynniki regresji $\mathbb{b}$ są realizacjami zmiennych losowych. Odpowiadają współczynnikom w równaniu obiektu, ale nie muszą być im równe. Zgodnie z równaniem (13.25) współczynniki zależą liniowo od zmiennej losowej $\mathbb{y}$, więc też są zmiennymi losowymi o rozkładzie normalnym, zaś ich wartość oczekiwana $\mathrm{E}(\mathbb{b}) = \beta$. Zatem można wyznaczyć macierz kowariancji wektora $\mathbb{b}$, która ma postać

$$\mathrm{cov}(\mathbb{b}) = \mathrm{E}\left((\mathbb{b} - \mathrm{E}(\mathbb{b}))(\mathbb{b} - \mathrm{E}(\mathbb{b}))^T\right) = \mathrm{E}\left((\mathbb{b} - \beta)(\mathbb{b} - \beta)^T\right)$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Ponieważ

$$\mathbb{b} - \beta = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{y} - \beta = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T(\mathbb{X}\beta + \varepsilon) - \beta = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\varepsilon$$

to po uwzględnieniu, że macierz $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ jest symetryczna, macierz kowariancji wektora $\mathbb{b}$ można wyznaczyć jako

$$\begin{aligned}\operatorname{cov}(\mathbb{b}) &= \mathrm{E}\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\varepsilon\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\varepsilon\right)^T\right) = \\ &= \mathrm{E}\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\varepsilon\varepsilon^T\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1T}\right) = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathrm{E}\left(\varepsilon\varepsilon^T\right)\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1} = \\ &= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{I}\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\sigma^2 = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{X}\mathbb{I}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\sigma^2 = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\sigma^2\end{aligned} \tag{13.28}$$

W rozwiniętym zapisie, powyższa zależność przyjmie postać

$$\operatorname{cov}(\mathbb{b}) = \begin{bmatrix} \operatorname{var}(b_0) & \operatorname{cov}(b_0b_1) & \dots & \operatorname{cov}(b_0b_K) \\ \operatorname{cov}(b_1b_0) & \operatorname{var}(b_1) & \dots & \operatorname{cov}(b_1b_K) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(b_Kb_0) & \operatorname{cov}(b_Kb_1) & \dots & \operatorname{var}(b_K) \end{bmatrix} = \begin{bmatrix} c_{00} & c_{01} & \dots & c_{0K} \\ c_{10} & c_{11} & \dots & c_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{K0} & c_{K1} & \dots & c_{KK} \end{bmatrix}\sigma^2 \tag{13.29}$$

gdzie c_{ij} są elementami macierzy odwrotnej $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ i są zwane mnożnikami Gaussa, zaś σ^2, jak już wcześniej wspomniano, jest wariancją błędu. Z powyższej zależności wynikają wzory na wariancję współczynników b_i i kowariancję współczynników b_ib_j. Współczynniki korelacji między współczynnikami b_i oraz b_j można wyliczyć z zależności

$$R(b_i, b_j) = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}$$

Elementy diagonalne macierzy $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ charakteryzują wariancję współczynników regresji b_i, a pozostałe jej elementy charakteryzują kowariancję współczynników b_i oraz b_j. Z tego powodu macierz $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ nazywa się macierzą kowariancyjną. Trzeba podkreślić, iż zależność (13.29) wskazuje, że współczynniki regresji nie są wyznaczane niezależnie od siebie. Ponieważ macierz $\mathbb{X}$ zależy od przyjętych wartości wejściowych w eksperymencie, można zbudować taki plan eksperymentu, w którym macierz $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ będzie macierzą diagonalną, zaś współczynniki regresji b_i będą wyznaczane niezależnie od siebie.

Istotność współczynników regresji. Współczynniki regresji b_i są zmiennymi losowymi. Jeżeli zakłócenia mają rozkład normalny, to współczynniki regresji też mają rozkład normalny. Wartością oczekiwaną współczynników b_i jest β_i, zaś odchylenie standardowe wynosi $\sigma_{b_i} = \sigma\sqrt{c_{ii}}$. Zatem unormowana zmienna losowa

$$U = \frac{b_i - \beta_i}{\sigma_{b_i}} = \frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

ma standardowy rozkład normalny $N(0,1)$. Ponieważ nie jest znana wartość odchylenia standardowego σ, konieczne jest jej estymowanie. Wygodnym nieobciążonym estymatorem odchylenia standardowego σ jest wyrażenie

$$\hat{\sigma}^2 = \frac{1}{N-K-1}\sum_{i=1}^{N}(y_i - \hat{y}_i(b_0, b_1, \ldots, b_K))^2 = \hat{\sigma}^2_{y-\hat{y}} \tag{13.30}$$

które jest jednocześnie wariancją resztową $\sigma^2_{y-\hat{y}}$ zmiennej losowej $y-\hat{y}$ i ma rozkład χ^2 o $N-K-1$ stopniach swobody. Zatem zamiast zmiennej losowej U można wyznaczyć zmienną losową t

$$t = \frac{b_i - \beta_i}{\hat{\sigma}_{y-\hat{y}}\sqrt{c_{ii}}}$$

która ma rozkład t-Studenta o $N-K-1$ stopniach swobody. Przyjmując poziom istotności α, można skonstruować przedział ufności $1-\alpha$ i znaleźć taką wartość rozkładu t-Studenta o $N-K-1$ stopniach swobody, że spełniona jest równość

$$P(|t_{N-K-1}| < t_{kr}) = 1 - \alpha$$

Można teraz wyznaczyć przedział ufności dla współczynników regresji w postaci układu nierówności

$$b_i - t_{kr}\hat{\sigma}_{y-\hat{y}}\sqrt{c_{ii}} < \beta_i < b_i + t_{kr}\hat{\sigma}_{y-\hat{y}}\sqrt{c_{ii}} \tag{13.31}$$

określający przedział ufności dla współczynnika regresji β_i na poziomie ufności $1-\alpha$. Uzyskaną zmienną można też użyć do testowania hipotez dotyczących współczynników regresji.

Hipoteza o braku zależności liniowej zakłada, że nie ma zależności liniowej między wyjściem obiektu y a wejściem $f_i(\mathbb{x})$, czyli $\beta_i = 0$. Decyduje wartość statystyki

$$t_{hip} = \frac{|b_i|}{\hat{\sigma}_{y-\hat{y}}\sqrt{c_{ii}}}$$

Po przyjęciu poziomu istotności α i liczby stopni swobody $N-K-1$, wyznacza się krytyczną wartość z rozkładu t-Studenta

$$P(|t_{N-K-1}| < t_{kr}) = 1 - \alpha$$

Jeżeli obliczona wartość $t_{hip} > t_{kr}$, to należy odrzucić hipotezę o nieistotności współczynnika regresji. W przypadku przeciwnym nie ma podstaw do odrzucenia hipotezy.

Hipoteza o wartości współczynnika regresji pozwala sprawdzić, czy można przyjąć, że wartość współczynnika β_i wynosi β_i^0, czyli $\beta_i = \beta_i^0$. Wartość statystyki wynikającą z hipotezy wyznacza się z zależności

$$t_{hip} = \frac{b_i - \beta_i^0}{\hat{\sigma}_{y-\hat{y}}\sqrt{c_{ii}}}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po przyjęciu poziomu istotności α i liczby stopni swobody $N-K-1$, wyznacza się krytyczną wartość z rozkładu t-Studenta

$$P(|t_{N-K-1}| \geqslant t_{kr}) = \alpha$$

Gdy $t_{hip} > t_{kr}$, hipotezę można odrzucić, w przeciwnym przypadku nie ma podstaw do odrzucenia hipotezy.

13.5.3. Przedziały ufności dla funkcji regresji

Można je wyznaczyć na podstawie macierzy obserwacji wejść $\mathbb{X}$ oraz wektora obserwacji wyjść y. W niniejszym podrozdziale zostanie wyznaczony przedział ufności na poziomie ufności $1-\alpha$ dla prognozowanej wartości wyjścia $\hat{y}^0$ odpowiadającej wartościom wejść $x_1^0, x_2^0, \ldots, x_K^0$, zgodnie we wzorem (13.15), w skrócie zapisanym jako

$$\hat{y}^0 = \left(\mathbb{X}^0\right)^T \mathbb{b}$$

gdzie $(\mathbb{X}^0)^T = [1, f_1(\mathbb{x}^0), \ldots, f_K(\mathbb{x}^0)]$. Ponieważ $\mathbb{b}$ jest zmienną losową o rozkładzie normalnym, jeżeli zakłócenia są zmienną losową o rozkładzie normalnym, to $\hat{y}^0$ też jest zmienną losową o rozkładzie normalnym i wartości oczekiwanej

$$\mathrm{E}\left(\hat{y}^0\right) = \mathrm{E}\left(\left(\mathbb{X}^0\right)^T \mathbb{b}\right) = \left(\mathbb{X}^0\right)^T \mathrm{E}(\mathbb{b}) = \left(\mathbb{X}^0\right)^T \beta$$

i wariancji równej

$$\begin{aligned}
\mathrm{Var}(\hat{y}^0) &= \mathrm{E}\left(\left(\hat{y}^0 - \mathrm{E}(\hat{y}^0)\right)\left(\hat{y}^0 - \mathrm{E}(\hat{y}^0)\right)^T\right) = \\
&= \mathrm{E}\left(\left(\left(\mathbb{X}^0\right)^T \mathbb{b} - \left(\mathbb{X}^0\right)^T \beta\right)\left(\left(\mathbb{X}^0\right)^T \mathbb{b} - \left(\mathbb{X}^0\right)^T \beta\right)^T\right) = \\
&= \mathrm{E}\left(\left(\left(\mathbb{X}^0\right)^T (\mathbb{b} - \beta)\right)\left(\left(\mathbb{X}^0\right)^T (\mathbb{b} - \beta)\right)^T\right) = \mathrm{E}\left(\left(\mathbb{X}^0\right)^T (\mathbb{b} - \beta)(\mathbb{b} - \beta)^T \mathbb{X}^0\right) = \\
&= \left(\mathbb{X}^0\right)^T \underbrace{\mathrm{E}\left((\mathbb{b} - \beta)(\mathbb{b} - \beta)^T\right)}_{\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\sigma^2} \mathbb{X}^0 = \left(\mathbb{X}^0\right)^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} \mathbb{X}^0 \sigma^2 \qquad (13.32)
\end{aligned}$$

Jak widać z powyższej zależności, wariancja wartości wyjściowej $\hat{y}^0$ zależy nie tylko od macierzy kowariancji współczynników $\mathbb{b}$, ale także od wartości wektora wejść $\mathbb{x}^0$. Co powoduje, że ta wariancja nie jest stała. Ponieważ nie jest znana wartość σ^2, można wykorzystać jej estymator $\hat{\sigma}^2_{y-\hat{y}}$, zgodnie ze wzorem (13.30). Przywołaną zależność (13.30) można uprościć do [34]

$$\hat{\sigma}^2_{y-\hat{y}} = \frac{1}{N-K-1}\left(\mathrm{y}^T\mathrm{y} - \mathbb{b}^T\mathbb{X}^T\mathrm{y}\right)$$

Zatem przyjmując poziom ufności $1-\alpha$ oraz określając dla niego wartość krytyczną rozkładu t-Studenta t_{kr} o liczbie stopni swobody $N-K-1$, można zapisać przedział ufności dla prognozowanej wartości wyjścia $\hat{y}^0$ jako

$$\hat{y}^0 \pm t_{kr}\hat{\sigma}^2_{y-\hat{y}}\sqrt{\left(\mathbb{X}^0\right)^T \left(\mathbb{X}^T\mathbb{X}\right)^{-1} \mathbb{X}^0}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

W ten sposób można wyznaczyć zakres wartości wyjściowych wynikających z modelu stochastycznego.

W praktyce jednakże metoda analizy regresji często nie pozwala na uzyskanie dobrych wyników. Taki stan rzeczy najczęściej jest spowodowany wadą samego eksperymentu. Okazuje się też, że wpływ zakłóceń na wielkość wyjściową jest często dużo większy niż wpływ wielkości wejściowych. Powoduje to, że zmienność stanu wyjścia od stanów wejść nie jest istotna statystycznie. Ten stan rzeczy skutkuje koniecznością wybierania w doświadczeniach ekstremalnych wartości wejściowych, aby zminimalizować wpływ zakłóceń. Najważniejszą informacją, którą czytelnik koniecznie powinien sobie przyswoić, jest to, że istotność otrzymanej funkcji regresji świadczy jedynie o korelacji pomiędzy wielkościami, lecz nie świadczy o istnieniu związku przyczynowo-skutkowego pomiędzy wielkością wyjściową a wielkościami wejściowymi.

13.5.4. Adekwatność funkcji regresji

Podstawowe pytanie, które pojawia się przy modelowaniu obiektów rzeczywistych za pomocą arbitralnie przyjętych funkcji, na przykład wielomianów, dotyczy adekwatności przyjętego modelu. Innymi słowy, czy funkcja, której współczynników się poszukuje, odpowiada równaniu, które opisuje zachowanie obiektu.

Jeżeli budowany model matematyczny oparty jest na znajomości zasad fizyki i na tej podstawie opisuje się układ, zaś regresję wykorzystuje się do identyfikacji wartości współczynników modelu, to można oczekiwać takiej adekwatności. Kłopoty pojawiają się wtedy, gdy analizuje się wpływ przyjętych uproszczeń na uzyskaną ogólną postać modelu matematycznego obiektu. Tym zagadnieniem zajmuje się wielu autorów, szeroko opisując ogólne i szczegółowe zasady stosowania uproszczeń w modelowaniu.

W niniejszym podrozdziale, na przykładzie funkcji regresji pierwszego stopnia z interakcjami między czynnikami, zostanie przeanalizowany wpływ pomijania niektórych składników w funkcji regresji. Omawiany przykład będzie mieć istotne znaczenie przy omawianiu ułamkowych planów eksperymentu.

Zakłada się, że model zapisany równaniem (13.15) dobrze opisuje analizowany obiekt przy uwzględnieniu K składników. W takim przypadku współczynniki b_i są nieobciążonymi estymatorami nieznanych parametrów β_i z równania (13.16).

Natomiast w przypadku, gdy do opisu obiektu użyje się $M < K$ wyrazów, czyli przyjęty zostanie model następujący

$$\hat{y} = b_0 + b_1 f_1(\mathbb{x}) + b_2 f_2(\mathbb{x}) + \ldots + b_M f_M(\mathbb{x}) = b_0 + \sum_{i=1}^{M} b_i f_i(\mathbb{x}) \tag{13.33}$$

to wyznaczone w tych dwóch modelach wartości współczynników b_i muszą się różnić.

Równanie opisujące wartość oczekiwaną wyjścia obiektu ma postać (13.34)

$$\mathrm{E}(y) = \beta_0 + \sum_{i=1}^{K} \beta_i f_i(\mathbb{x}) \tag{13.34}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Po wprowadzeniu oznaczeń $\tilde{f}_i = f_{M+i}$ oraz $\tilde{\beta}_i = \beta_{M+i}$ można zapisać równanie układu (13.34) w postaci (13.35), co pozwoli na oszacowanie różnicy pomiędzy dwoma omawianymi modelami.

$$\mathrm{E}(y) = \beta_0 + \sum_{i=1}^{M} \beta_i f_i(\mathbb{x}) + \underbrace{\sum_{i=1}^{K-M} \tilde{\beta}_i \tilde{f}_i(\mathbb{x})}_{\text{obciążenie modelu}} \tag{13.35}$$

Druga suma w równaniu (13.35) stanowi obciążenie modelu.

Po wprowadzeniu oznaczeń pozwalających na zapis wektorowy powyższych równań uzyskuje się zależność macierzową

$$\hat{\mathbb{y}} = \mathbb{X}\mathbb{b}$$

gdzie $\mathbb{X}$ jest macierzą przedstawioną w równaniu (13.17). Po wprowadzeniu analogicznego oznaczenia $\tilde{\mathbb{X}}$

$$\tilde{\mathbb{X}} = \begin{bmatrix} \tilde{f}_1(\mathbb{x}_1) & \tilde{f}_2(\mathbb{x}_1) & \dots & \tilde{f}_{K-M}(\mathbb{x}_1) \\ \tilde{f}_1(\mathbb{x}_2) & \tilde{f}_2(\mathbb{x}_2) & \dots & \tilde{f}_{K-M}(\mathbb{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{f}_1(\mathbb{x}_N) & \tilde{f}_2(\mathbb{x}_N) & \dots & \tilde{f}_{K-M}(\mathbb{x}_N) \end{bmatrix}$$

oraz $\tilde{\beta} = [\beta_1, \beta_2, \dots, \beta_{K-M}]^T$ wartość oczekiwana wyjścia z obiektu (13.35) może być zapisana w postaci macierzowej

$$\mathrm{E}(\mathbb{y}) = \mathbb{X}\beta + \tilde{\mathbb{X}}\tilde{\beta}$$

Macierz $\tilde{\mathbb{X}}$ oraz wektor $\tilde{\beta}$ służą do oceny wpływu pominiętych współczynników, które powinny zostać uwzględnione w modelu. Po przeprowadzeniu postępowania związanego z minimalizacją sumy kwadratów odchyłek wyjścia modelu $\hat{\mathbb{y}}$ i wartości oczekiwanej wyjścia obiektu $\mathrm{E}(\mathbb{y})$ uzyska się wyrażenie na oszacowanie wartości współczynników

$$\begin{aligned} \mathrm{E}(\mathbb{b}) &= \mathrm{E}\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{y}\right] = \mathrm{E}\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\left(\mathbb{X}\beta + \tilde{\mathbb{X}}\tilde{\beta}\right)\right] = \\ &= \mathrm{E}\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{X}\beta\right] + \mathrm{E}\left[\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}}\tilde{\beta}\right] = \beta + \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}}\tilde{\beta} \end{aligned} \tag{13.36}$$

Powyższy wynik wskazuje, że estymator $\mathbb{B}$ jest estymatorem obciążonym współczynników regresji β, a obciążenie jest równe $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}}\tilde{\beta}$. Macierz $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}}$ nazywa się macierzą obciążeń. Wartość oczekiwana współczynników b_i nie jest równa β_i, ale

$$\mathrm{E}(b_i) = \beta_i + \sum_{j=1}^{K-M} a_{ij}\tilde{\beta}_j \tag{13.37}$$

gdzie a_{ij} są elementami macierzy obciążeń. Innymi słowy, wyestymowane współczynniki są pewnymi kombinacjami liniowymi współczynników β_i oraz $\tilde{\beta}_j$, przy czym $j = 1, 2, \dots, K - M$ [49].

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

13.6. Podsumowanie

W niniejszym rozdziale skrótowo omówiono pewne podstawowe koncepcje rachunku prawdopodobieństwa i statystyki matematycznej. Wiadomości dotyczące regresji, miar statystycznych czy testowania hipotez są szczególnie przydatne przy realizowaniu eksperymentów na obiektach fizycznych. W większości przypadków, eksperymenty numeryczne są pozbawione elementów losowych. Mimo to wiadomości dotyczące statystyki będą przydatne w rozdziale dotyczącym planowania eksperymentów.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rozdział 14

Planowanie eksperymentu

14.1. Wstęp

Planowanie eksperymentu ma długą historię. Początki sięgają XVI wieku, gdy rozwijająca się nauka potrzebowała zmniejszenia liczby kosztownych eksperymentów bez obniżania wiarygodności uzyskiwanych wyników. Początkowo systematyczne badania prowadzone były metodą zmiany wartości pojedynczych czynników. Prowadziło to do dużej liczby doświadczeń oraz, niestety, utrudniało wykrycie interakcji pomiędzy czynnikami wpływającymi na wartości wyjściowe obiektu, a w konsekwencji do wielokrotnych doświadczeń przy różnych wartościach poszczególnych wejść (czynników).

Trudności te były przyczyną powstania doświadczeń czynnikowych, a w konsekwencji metody planowania eksperymentu. Zagadnienie planowania eksperymentu przedstawił po raz pierwszy Ronald Aylmer Fischer w połowie lat trzydziestych ubiegłego wieku, tworząc podstawy analizy wariancyjnej. W 1945 roku David John Finney zaproponował ułamkowe plany eksperymentów czynnikowych. Praca zespołu autorów George'a Edwarda Pelhama Boxa oraz K. B. Wilsona z 1951 r., zatytułowana: *On the experimental attainement of optimum conditions* [6], stanowiła przełom. Opisana metoda eksperymentalnej optymalizacji obiektów była intuicyjna i bazowała na metodzie planowania eksperymentu. To właśnie ona spowodowała dalszy rozwój teorii i przyczyniła się do szerokiego rozpropagowania planowania eksperymentu jako podstawowej metody badawczej. Opisana przez nich metoda w literaturze jest spotykana pod nazwą metody Boxa–Wilsona.

Planowanie eksperymentu obejmuje wiele zagadnień. Obecnie rozróżnia się dwa obszary zastosowania tych metod. Po pierwsze, jest ono stosowane podczas eksperymentowania na obiektach fizycznych, ale stosuje się je też podczas eksperymentów numerycznych. Funkcjonują też dwa określenia związane z planowaniem eksperymentu. DoE, angielski skrót określenia ***D**esign **of** **E**xperiments*, oznacza planowanie eksperymentu i zasadniczo dotyczy eksperymentów na obiektach rzeczywistych, gdzie konieczne jest branie pod uwagę nieznanych losowych zakłóceń oraz związanych z nimi zależności stochastycznych.

Z kolei skrót DACE rozwija się jako ***D**esign/**A**nalysis of **C**omputer **E**xperiments*. Ponieważ obecnie buduje się komputery, których działanie jest deterministyczne, więc wyniki obliczeń zazwyczaj są identyczne. Daje to możliwość pewnego uproszczenia procesu opracowania wyników.

Chociaż aparat matematyczny oraz modele obiektów są podobne, to różnią się powtarzalnością eksperymentów komputerowych. Jedną z konsekwencji tej różnicy

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

jest sposób wybierania punktów w planie eksperymentu. O ile w tradycyjnych eksperymentach na obiektach rzeczywistych należy dołożyć starań, by wartości podawane na wejścia układu były możliwie odległe od siebie, o tyle w eksperymentach komputerowych nie ma takiej konieczności. W związku z tym w fizycznych eksperymentach często wybiera się wierzchołki planu eksperymentu, by zmaksymalizować różnicę odpowiedzi układu i zmniejszyć względny wpływ zakłóceń. Zaś w eksperymentach komputerowych można wybierać wartości wymuszeń znacznie lepiej pokrywające całą przestrzeń pracy układu.

W eksperymentach komputerowych najczęściej chodzi o takie wybranie wartości wejściowych z całej przestrzeni wartości dopuszczalnych, by uzyskać wiarygodną informację o odpowiedzi układu przy minimalnej liczbie wykonanych doświadczeń.

Od dawna planowanie eksperymentu było wykorzystywane do budowania modeli empirycznych. Jak wspomniano we wstępie do niniejszego rozdziału, rozwój numerycznych metod rozwiązywania problemów technicznych spowodował swego rodzaju renesans tych modeli. Lecz zamiast do opisu rzeczywistości stosuje się je do opisu dyskretnych modeli numerycznych. W ten sposób uzyskuje się ciągłe modele analityczne opisujące zachowanie modeli numerycznych. W praktyce daje to kolosalne korzyści, ponieważ można wykorzystywać dobrze rozwinięty aparat matematyczny dotyczący badania funkcji ciągłych i na przykład dokonywać optymalizacji stosunkowo małym nakładem czasu.

Powierzchnie drugiego stopnia są wygodnymi modelami matematycznymi (modelami zastępczymi) obiektów rzeczywistych wykorzystywanymi między innymi właśnie na potrzeby optymalizacji. Ich główną zaletą jest możliwość odwzorowania ekstremów lokalnej charakterystyki obiektu. Do wyznaczenia parametrów kwadratowej powierzchni odpowiedzi konieczne jest wykorzystanie planów, w których zmiany parametrów wejściowych odbywają się przynajmniej na trzech poziomach. Pełne plany trójpoziomowe wymagają znacznie więcej doświadczeń niż wymagane do efektywnego wyznaczenia tych współczynników.

Pojęcie powierzchnia odpowiedzi nie ogranicza się tylko do powierzchni drugiego stopnia, a do ich budowania wykorzystuje się nie tylko wielomiany. Wspomniane powyżej plany eksperymentu, które pozwalają wyznaczać powierzchnie odpowiedzi zostaną opisane w dalszej części niniejszego rozdziału. W tym miejscu zostaną tylko zasygnalizowane pewne kryteria, które służą do oceny planów. Wymagania stawiane planom eksperymentu pozwalającym wyznaczyć powierzchnię odpowiedzi ze składnikami kwadratowymi można ująć następująco:

— zadowalający rozkład informacji w badanym obszarze połączony z niezależnością od wyboru układu współrzędnych (rotatabilność);
— wartości powierzchni odpowiedzi są możliwie bliskie wartościom z pomiarów, czyli błąd predykcji jest minimalizowany;
— dobre wykrywanie braku dopasowania;
— estymacja błędów;
— sprawdzenie stałej wariancji;
— estymacja transformacji;
— wygodne blokowanie doświadczeń;

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— sekwencyjna budowa planów wyższych rzędów z planów podstawowych;
— możliwie mała liczba doświadczeń;
— wygodna analiza i interpretacja wyników (najlepiej graficzna);
— mała wrażliwość na błędy w ustawianiu wartości wejściowych.

Powyższe kryteria stanowią jednocześnie kryteria oceny planów eksperymentu.

Po tym krótkim wstępie zostaną omówione podstawowe plany dwupoziomowe zarówno całkowite, jak i ułamkowe, a następnie plany trójpoziomowe i wielopoziomowe.

14.2. Plany dwupoziomowe

Najpopularniejszymi planami eksperymentu są plany dwupoziomowe. Opisują one serie doświadczeń przeprowadzane na dwóch poziomach, co znaczy, że wartości wejść modelu przedstawionego na rysunku 13.3 przyjmują wartości ekstremalne. Pod pojęciem wartości ekstremalnych rozumie się maksymalną i minimalną wartość z zakresu zmienności danego wejścia. Symbolicznie w macierzy eksperymentu zapisuje się to przez $+1$ dla wartości maksymalnej oraz -1 dla wartości minimalnej. Niektóre publikacje w zapisie opuszczają liczbę 1 i podają jedynie znaki $+$ oraz $-$, odpowiednio. Jeszcze inni autorzy używają nieco mylącej notacji 1 oraz 0. Nie zmienia to faktu, że zmienne wejściowe przyjmują dwa poziomy wartości, stąd też nazwa tej rodziny planów eksperymentu.

Pełne doświadczenie czynnikowe na dwóch poziomach rozpatruje takie zestawy wartości wejść, w których występują wszelkie możliwe wektory, które mogą zostać zbudowane przez przypisywanie każdemu czynnikowi wartości ± 1. Dla S czynników liczba takich wektorów będzie wynosić $N = 2^S$. Rozpatrywany plan eksperymentu podaje informację, który z czynników występuje na poziomie górnym $+1$, a który na poziomie dolnym -1.

Dlaczego przyjmuje się tylko dwie wartości dla każdego wejścia? Można wymienić przynajmniej kilka dobrych powodów, z których najistotniejsze to:

— możliwość oceny zakresu zmienności wyjścia dla całego zakresu wejściowego,
— możliwość oceny wrażliwości zmian wartości wyjściowej na zmiany wartości poszczególnych wejść,
— ich prostota, a co za tym idzie, ekonomiczność,
— zdobycie informacji pomagającej zaplanować eksperymenty trójpoziomowe i wielopoziomowe (jeśli takie są potrzebne).

Należy mieć na uwadze, że chociaż plan jest dwupoziomowy, to warto w eksperymencie uwzględnić wartość środkową, czyli leżącą w środku analizowanego obszaru zmian wartości wejść.

Aby zilustrować te rozważania, poniżej została zamieszczona tabela 14.1 przedstawiająca prosty całkowity dwupoziomowy plan eksperymentu dla dwóch czynników. W ostatniej kolumnie znajduje się kod zgodny z systemem kodowania zaproponowanym przez Franka Yatesa, polegający na tym, że poszczególnym czynnikom przyporządkowuje się kolejne litery alfabetu. Poszczególne kombinacje poziomów w doświad-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

czeniach podaje się, zapisując te czynniki, które przyjmują wartości górne (maksymalne). I tak w doświadczeniu numer dwa czynnik x_1 przyjmuje wartość maksymalną, zatem zapis kodowy wiersza sprowadza się do litery a. Przy doświadczeniu czwartym oba czynniki mają wartość maksymalną, wobec tego zapis kodowy to ab. W doświadczeniu pierwszym, gdy nie ma czynników występujących na górnym poziomie, wiersz koduje się jako (1).

Tabela 14.1. Plan eksperymentu typu 2^S dla dwóch czynników

Doświadczenie	Czynnik 1 (x_1)	Czynnik 2 (x_2)	Pkt	Kod
1	-1	-1	p_1	(1)
2	$+1$	-1	p_2	a
3	-1	$+1$	p_3	b
4	$+1$	$+1$	p_4	ab

Plan przedstawiony w tabeli 14.1 pozwala na wyznaczenie czterech współczynników funkcji regresji w postaci

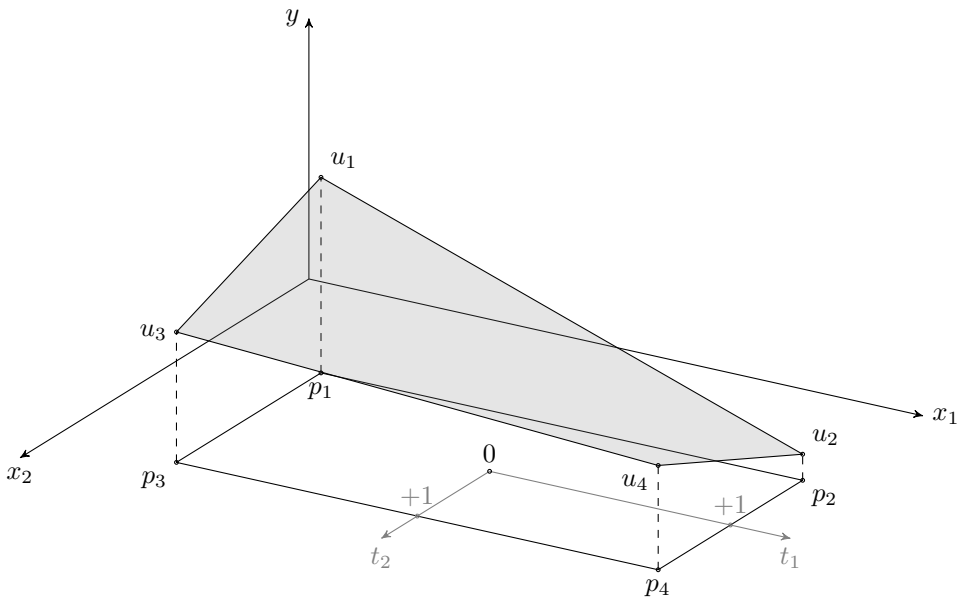
$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$$

Dodanie kolejnego czynnika, czyli kolejnego wejścia do obiektu z rysunku 13.3, wymaga dodania kolejnej kolumny do macierzy eksperymentu. Graficzną ilustrację opisywanego postępowania stanowi rysunek 14.1, na którym przedstawiono zmiany wartości wejść x_1 oraz x_2 we współrzędnych naturalnych oraz zmiany wartości wejść we współrzędnych lokalnych x_1^0 oraz x_2^0. Prostokąt leżący na płaszczyźnie $x_1 0 x_2$ przedstawia analizowany obszar zmian wejść. Naroża prostokąta p_1, p_2, p_3 oraz p_4 stanowią graficzną reprezentację kolejnych wierszy powyższego planu eksperymentu. Wartości wyjścia dla poszczególnych doświadczeń opisują punkty u_1, u_2, u_3 oraz u_4. Szary obszar stanowi powierzchnię odpowiedzi badanego układu.

Używanie wartości ± 1 dla czynników nazywane jest standaryzowaniem zmiennych wejściowych. Taka standaryzacja powoduje, że macierz eksperymentu ma korzystne własności. Na przykład kolumny planu eksperymentu są ortogonalne, co oznacza, że iloczyny skalarne poszczególnych kolumn macierzy są zerami. Standaryzacja zmiennych wejściowych przeprowadzana jest w taki sposób, by w całym zakresie zmienności parametrów wejściowych przyjmowały one wartości od -1 do 1. Przeprowadza się ją następująco: niech ${x_i}^0$ oznacza środek przedziału zmian dla i-tej zmiennej, wektor $[x_1^0, x_2^0, \ldots, x_S^0]^T$ stanowi punkt, w którego otoczeniu przeprowadza się eksperyment. Cała standaryzacja przebiega zgodnie z poniższymi zależnościami, zilustrowanymi rysunkiem 14.2.

$$x_i^0 = \frac{x_i^{max} + x_i^{min}}{2}, \quad \Delta x_i = \frac{x_i^{max} - x_i^{min}}{2} \tag{14.1}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 14.1. Schemat planu eksperymentu czynnikowego na dwóch poziomach dla dwóch zmiennych x_1 i x_2

Wartości zmiennej standaryzowanej wylicza się według wzoru

$$t_i = \frac{x_i - x_i^0}{\Delta x_i} \tag{14.2}$$

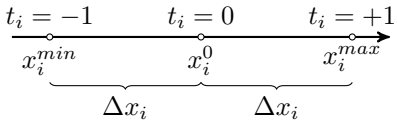
i używa się ich do zapisu macierzy eksperymentu. Warto zdawać sobie sprawę z konsekwencji standaryzacji zmiennych wejściowych. W przypadku zmiennych wyrażonych w jednostkach naturalnych, jak pokazano na rysunku 14.1, punkty planu stanowią wierzchołki prostokąta definiującego obszar badań, ograniczonego wartościami maksymalną oraz minimalną. Przez wprowadzenie zmiennej standaryzowanej t_i środek układu współrzędnych przesuwany jest do punktu centralnego. Jednocześnie skala osi jest modyfikowana tak, że uzyskuje się kwadrat o wierzchołkach w punktach $(-1, -1)$, $(+1, -1)$, $(-1, +1)$ oraz $(+1, +1)$, jak pokazano na rysunku 14.1 oraz w tabeli 14.1. Można oczekiwać, że powinna istnieć możliwość powiązania wartości współczynników regresji dla zmiennych naturalnych i dla zmiennych standaryzowanych. Przy funkcjach regresji stopnia pierwszego (równanie (14.4)), ta zależność przyjmuje prostą postać, ponieważ ze wzoru (14.2) wynika, że istnieje jednoznaczna zależność pozwalająca przejść od zmiennych standaryzowanych do zmiennych naturalnych. Dla i-tej zmiennej wejściowej można ją zapisać w następującej postaci

$$x_i = x_i^0 + t_i \Delta x_i \tag{14.3}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$x_i^0 = \frac{x_i^{max} + x_i^{min}}{2}, \quad \Delta x_i = \frac{x_i^{max} - x_i^{min}}{2}$$

$$t_i = \frac{x_i - x_i^0}{\Delta x_i}$$



Rysunek 14.2. Standaryzacja zmiennych wejściowych

Można zatem równanie regresji postaci

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_S x_S = b_0 + \sum_{i=1}^{S} b_i x_i \tag{14.4}$$

zapisać za pomocą związku (14.3) w postaci

$$y = b_0 + \sum_{i=1}^{S} b_i x_i = b_0 + \sum_{i=1}^{S} b_i ({x_i}^0 + t_i \Delta x_i) = b_0 + \sum_{i=1}^{S} b_i {x_i}^0 + \sum_{i=1}^{S} b_i \Delta x_i t_i \tag{14.5}$$

dzięki czemu po niewielkich przekształceniach otrzymuje się równanie regresji liniowej dla zmiennych standaryzowanych

$$y = a_0 + \sum_{i=1}^{S} a_i t_i \tag{14.6}$$

gdzie współczynnik

$$a_0 = b_0 + \sum_{i=1}^{S} b_i x_i^0 \tag{14.7}$$

Natomiast współczynniki a_i wyznacza się z zależności

$$a_i = b_i \Delta x_i, \quad \text{dla } i = 1, 2, \ldots, S \tag{14.8}$$

W ten sposób wykazano związek między współczynnikami regresji przy zmiennych wejściowych standaryzowanych oraz przy zmiennych wejściowych w naturalnych jednostkach. Dla nieliniowych funkcji regresji uzyskane zależności nieco się komplikują.

Po przeprowadzeniu standaryzacji zmiennych wejściowych należy zbudować macierz eksperymentu taką, jak pokazano wcześniej na przykład w tabeli 14.1. W zależności od tego, czy przeprowadza się doświadczenia dla każdej możliwej kombinacji

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

wartości wejść, czy też dla pewnego podzbioru tych kombinacji, uzyskuje się plan całkowity typu 2^S lub ułamkowy typu 2^{S-p}.

Warto zapamiętać, że plan eksperymentu dwupoziomowego wykorzystywany jest do identyfikacji modeli liniowych z interakcjami między zmiennymi. Ponieważ liczba doświadczeń wprost wynika z liczby zmiennych wejściowych S oraz liczby poziomów 2 i wynosi 2^S, plany całkowite tego typu nazywa się 2^S.

Plany całkowite buduje się w sposób sekwencyjny przez składanie planów niższego stopnia, jak to zostało pokazane w tabeli 14.2. Wykorzystując plan 2^{S-1} można zbudować plan 2^S, dodając jeszcze jedną zmienną wejściową t_S (kolejną kolumnę planu) na poziomie -1 dla wszystkich wierszy planu 2^{S-1}. Następnie należy powtórzyć wszystkie wiersze planu 2^{S-1} ze zmienną t_S o wartości -1.

Tabela 14.2. Całkowite plany eksperymentu czynnikowego typu 2^S dla $S = 1, 2, 3$

	2^1	2^2	2^3	
Nr	t_1	t_2	t_3	Kod
1	-1	-1	-1	(1)
2	$+1$	-1	-1	a
3	-1	$+1$	-1	b
4	$+1$	$+1$	-1	ab
5	-1	-1	$+1$	c
6	$+1$	-1	$+1$	ac
7	-1	$+1$	$+1$	bc
8	$+1$	$+1$	$+1$	abc

W ten sposób z planu dla jednego czynnika kodowanego jako [(1), a] powstał plan dla dwóch czynników [(1), a, b, ab]. Następnie analogicznie został zbudowany plan dla trzech czynników, który został zakodowany jako [(1), a, b, ab, c, ac, bc, abc], zaś plan dla czterech czynników typu 2^4 wyglądałby tak jak dwukrotne powtórzenie planu dla trzech czynników. Dla pierwszych 8 wierszy planu zmienna t_4 byłaby na dolnym poziomie -1, zaś dla następnych 8 wierszy byłaby na górnym poziomie $+1$.

W tabeli 14.3 przedstawiono macierz planowania i plan całkowitego eksperymentu czynnikowego na dwóch poziomach dla trzech czynników, zaś na rysunku 14.3 przedstawiono jego interpretację geometryczną. Plan taki wymaga wykonania ośmiu doświadczeń i pozwala na wyznaczenie współczynników funkcji regresji z ośmioma członami. Funkcja taka może mieć na przykład postać (14.9).

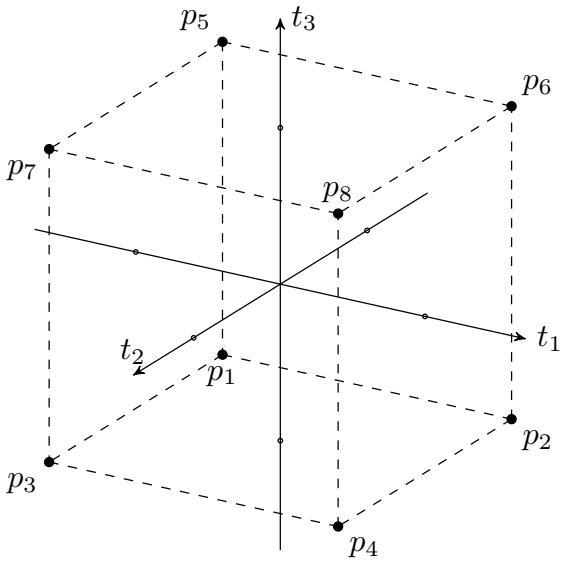
$$y = a_0 + a_1 t_1 + a_2 t_2 + a_3 t_3 + a_{12} t_1 t_2 + a_{13} t_1 t_3 + a_{23} t_2 t_3 + a_{123} t_1 t_2 t_3 \qquad (14.9)$$

Czasem na podstawie wiedzy o obiekcie można stwierdzić, że nie wszystkie interakcje między wejściami są istotne i niektóre można pominąć, bo praktycznie nie zachodzą. W takim przypadku bardzo pomocne są plany ułamkowe eksperymentu.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 14.3. Macierz planowania i plan pełnego eksperymentu czynnikowego typu 2^3

Nr	t_0	t_1	t_2	t_3	t_1t_2	t_1t_3	t_2t_3	$t_1t_2t_3$	Kod
1	+1	−1	−1	−1	+1	+1	+1	−1	(1)
2	+1	+1	−1	−1	−1	−1	+1	+1	a
3	+1	−1	+1	−1	−1	+1	−1	+1	b
4	+1	+1	+1	−1	+1	−1	−1	−1	ab
5	+1	−1	−1	+1	+1	−1	−1	+1	c
6	+1	+1	−1	+1	−1	+1	−1	−1	ac
7	+1	−1	+1	+1	−1	−1	+1	−1	bc
8	+1	+1	+1	+1	+1	+1	+1	+1	abc



Rysunek 14.3. Graficzna reprezentacja planu 2^3

Jak widać z tabeli 14.3 liczba doświadczeń dla trzech czynników wynosi cztery, i dla każdego następnego czynnika zostanie podwojona. Dlatego plany całkowite eksperymentów czynnikowych na dwóch poziomach raczej nie są stosowane dla większej niż cztery liczby zmiennych wejściowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

14.3. Plany dwupoziomowe ułamkowe

W przypadku większej liczby zmiennych wejściowych przeprowadzenie eksperymentu całkowitego może być praktycznie niemożliwe ze względu na dużą liczbę koniecznych do przeprowadzenia doświadczeń. W praktyce przyjmuje się, że graniczną liczbą zmiennych wejściowych, dla których opłacalne może być realizowanie eksperymentu całkowitego, jest 10. Rozwiązaniem, stanowiącym kompromis między kosztami a jakością wyników, są plany ułamkowe o 2^{S-P} doświadczeniach. Plany ułamkowe zawierają jedynie pewien podzbiór doświadczeń z planu całkowitego. Zagadnienie budowy planu ułamkowego nie jest trywialne i należy uwzględnić wiele czynników, które mają wpływ zarówno na koszty, jak i na jakość uzyskanych wyników. Plany ułamkowe tworzy się jako połówkowe, ćwiartkowe, a także ósemkowe w zależności od wartości zmiennej P, odpowiednio $P = 1$, 2 i 3.

Łatwo można sprawdzić, że plan eksperymentu całkowitego typu 2^S, zapisany za pomocą zmiennych standaryzowanych, spełnia następujące warunki:

— symetrii doświadczeń względem środka eksperymentu, co można zapisać jako

$$\sum_{i=1}^{N} t_{ij} = 0, \quad j = 1, 2, \ldots, S \tag{14.10}$$

— ortogonalności, polegającej na zerowaniu się wszystkich iloczynów skalarnych wektorów kolumnowych macierzy eksperymentu, co można zapisać jako

$$\sum_{n=1}^{N} t_{ni} t_{nj} = 0, \quad i, j = 1, 2, \ldots, S, \; i \neq j \tag{14.11}$$

— równości sum kwadratów elementów we wszystkich kolumnach macierzy eksperymentu

$$\sum_{i=1}^{N} t_{ij}^2 = N, \quad j = 1, 2, \ldots, S \tag{14.12}$$

Plany eksperymentów ułamkowych tworzy się w taki sposób, by spełniały wymienione powyżej warunki (14.10), (14.11) oraz (14.12), podobnie jak dla eksperymentu całkowitego. Spełnienie tych warunków jest możliwe przy wybieraniu połowy, jednej czwartej, czy jednej ósmej części doświadczeń planu eksperymentu pełnego. Budując plany ułamkowe, należy pamiętać, że macierz eksperymentu musi uwzględniać zmiany wszystkich zmiennych wejściowych. Na przykład dla planu typu 2^3 przedstawionego w tabeli 14.3 można zaproponować dwa plany połówkowe. Oczywiste jest, że za pomocą planu połówkowego nie można wyznaczyć współczynników funkcji regresji w postaci (14.9), toteż trzeba zmniejszyć liczbę składników w tym modelu matematycznym. Wobec tego za pomocą eksperymentu ułamkowego nie będzie można wyznaczyć pełnej postaci funkcji regresji. Zatem zachodzi sytuacja opisana w podrozdziale 13.5.4. Co więcej, każdy plan ułamkowy będzie miał inną zdolność rozdzielczą przy identyfikacji określonych współczynników regresji. Zdolność rozdzielcza planu eksperymentu

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

mówi o tym, w jaki sposób zostają zaburzone identyfikowane współczynniki regresji w stosunku do funkcji regresji wyznaczanej na podstawie planu całkowitego.

Na przykład obiekt opisany przy realizacji planu całkowitego typu 2^3 z tabeli 14.3 funkcją regresji w postaci (14.9) zostaje poddany procedurze identyfikacji planem połówkowym. Niektórzy badacze mogliby w pierwszej chwili wybrać plan kodowany jako [(1), a, b, c]. Cztery doświadczenia pozwolą na wyznaczenie funkcji regresji w postaci

$$y = b_0 + b_1 t_1 + b_2 t_2 + b_3 t_3$$

Warto sprawdzić, czy współczynniki b_0, b_1, b_2 i b_3 odpowiadają współczynnikom a_0, a_1, a_2 oraz a_3. Plan eksperymentu zakodowany powyżej można opisać następującą macierzą wejść

$$\mathbb{X} = \begin{array}{c} \begin{matrix} 1 & t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} +1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \end{array} \begin{matrix} (1) \\ a \\ b \\ c \end{matrix}$$

Widać, że w stosunku do pełnej funkcji regresji (14.9) brakuje członów

$$a_{12} t_1 t_2 + a_{13} t_1 t_3 + a_{23} t_2 t_3 + a_{123} t_1 t_2 t_3$$

które zostaną uwzględnione w macierzy $\tilde{\mathbb{X}}$, stanowiąc poszczególne jej kolumny

$$\tilde{\mathbb{X}} = \begin{array}{c} \begin{matrix} t_1 t_2 & t_1 t_3 & t_2 t_3 & t_1 t_2 t_3 \end{matrix} \\ \begin{bmatrix} +1 & +1 & +1 & -1 \\ -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \end{array}$$

Zatem macierz obciążeń $\left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \tilde{\mathbb{X}}$ będzie wynosić

$$\left(\mathbb{X}^T \mathbb{X}\right)^{-1} \mathbb{X}^T \tilde{\mathbb{X}} = \begin{array}{c} \begin{matrix} a_{12} & a_{13} & a_{23} & a_{123} \end{matrix} \\ \begin{bmatrix} -1 & -1 & -1 & 2 \\ -1 & -1 & 4{,}44 \cdot 10^{-16} & 1 \\ -1 & -5{,}55 \cdot 10^{-17} & -1 & 1 \\ -1{,}11 \cdot 10^{-16} & -1 & -1 & 1 \end{bmatrix} \end{array} \approx$$

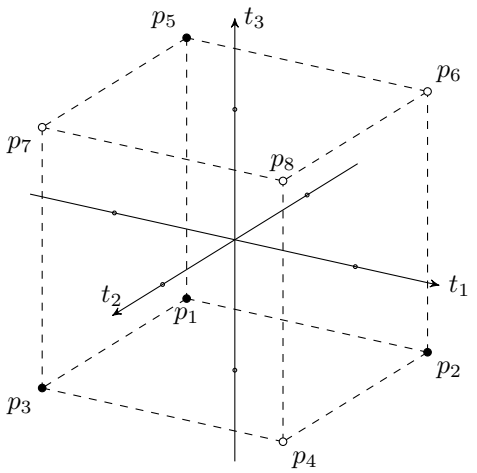
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

$$\approx \begin{array}{c} \begin{matrix} a_{12} & a_{13} & a_{23} & a_{123} \end{matrix} \\ \begin{bmatrix} -1 & -1 & -1 & 2 \\ -1 & -1 & 0 & 1 \\ -1 & 0 & -1 & 1 \\ 0 & -1 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{matrix} \tag{14.13}$$

Co, zgodnie z równaniem (13.37) oznacza, że wyznaczone współczynniki regresji b_i będą następującymi kombinacjami liniowymi współczynników a_i z równania (14.9)

$$\begin{aligned} b_0 &= a_0 - a_{12} - a_{13} - a_{23} + 2a_{123} \\ b_1 &= a_1 - a_{12} - a_{13} + a_{123} \\ b_2 &= a_2 - a_{12} - a_{23} + a_{123} \\ b_1 &= a_3 - a_{13} - a_{23} + a_{123} \end{aligned} \tag{14.14}$$

Uzyskany wynik świadczy o bardzo złej zdolności rozdzielczej wybranego planu ułamkowego. Dokładniejsza analiza macierzy wejść pozwala stwierdzić, że nie zostały spełnione wymagania stawiane planom ułamkowym. Nie jest spełnione wymaganie dotyczące symetrii doświadczeń wokół środka eksperymentu, opisane równaniem (14.10). Graficzną reprezentację tego planu przedstawia rysunek 14.4. Zaczernione punkty w narożach oznaczają przeprowadzane doświadczenia.



Rysunek 14.4. Plan ułamkowy [(1), a, b, c], zaczernione punkty reprezentują przeprowadzone doświadczenia

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Inny plan. Ponieważ rozważany plan eksperymentu nie ma zadowalających właściwości rozdzielczych, czyli wyestymowane współczynniki regresji są zaburzone w bardzo dużym stopniu, konieczne jest wybranie innego planu eksperymentu ułamkowego. Niech kodowanie planu zaczyna się $[a, b, c]$, czyli poszczególne wejścia przyjmują wartości maksymalne w poszczególnych doświadczeniach. Wobec tego pierwsze trzy wiersze planu wyglądają następująco

$$\mathbb{X} = \begin{array}{c} \begin{matrix} 1 & t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \end{array} \begin{matrix} \\ a \\ b \\ c \end{matrix}$$

W ostatnim wierszu planu w pierwszej kolumnie musi być $+1$, co wynika z postaci funkcji regresji (odpowiada wyrazowi wolnemu). Następne kolumny zostaną dobrane tak, by sumy poszczególnych kolumn się zerowały. Chodzi o to, by plan był symetryczny względem środka eksperymentu, co powoduje, że ostatni wiersz musi mieć postać $[+1\ +1\ +1\ +1]$. Kod ostatniego wiersza to abc. Zatem macierz eksperymentu połówkowego o znacznie lepszych właściwościach wygląda następująco

$$\mathbb{X} = \begin{array}{c} \begin{matrix} 1 & t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \\ +1 & +1 & +1 & +1 \end{bmatrix} \end{array} \begin{matrix} \\ a \\ b \\ c \\ abc \end{matrix}$$

Za pomocą przedstawionego powyżej sposobu zostanie wykonane sprawdzenie zdolności rozdzielczej opracowanego planu eksperymentu. Macierz $\tilde{\mathbb{X}}$ ma postać

$$\tilde{\mathbb{X}} = \begin{array}{c} \begin{matrix} a_{12} & a_{13} & a_{23} & a_{123} \end{matrix} \\ \begin{bmatrix} -1 & -1 & +1 & +1 \\ -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 \\ +1 & +1 & +1 & +1 \end{bmatrix} \end{array}$$

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

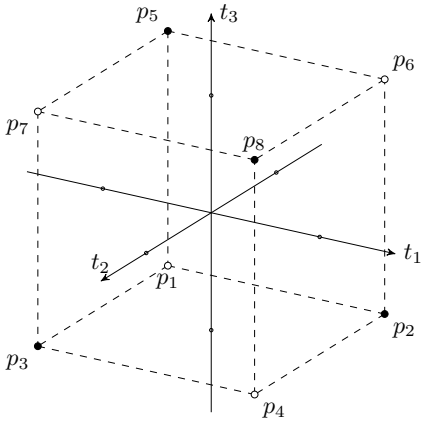
Zatem macierz $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}}$ jest postaci

$$\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\tilde{\mathbb{X}} = \begin{array}{c} \begin{matrix} a_{12} & a_{13} & a_{23} & a_{123} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{matrix}$$

co, zgodnie z równaniem (13.37), oznacza, że wyznaczone współczynniki regresji b_i będą następującymi kombinacjami liniowymi współczynników a_i z równania (14.9)

$$\begin{aligned} b_0 &= a_0 + a_{123} \\ b_1 &= a_1 + a_{23} \\ b_2 &= a_2 + a_{13} \\ b_1 &= a_3 + a_{12} \end{aligned} \tag{14.15}$$

Porównanie współczynników z układu (14.14) ze współczynnikami (14.15) jednoznacznie wskazuje, że drugi plan ma znacznie lepszą zdolność rozdzielczą, choć nie pozwala na wyznaczenie niezaburzonych współczynników regresji.



Rysunek 14.5. Plan ułamkowy [a, b, c, abc], zaczernione punkty reprezentują przeprowadzone doświadczenia

Opracowany plan eksperymentu przedstawiono graficznie na rysunku 14.5. Podobna sytuacja ma miejsce przy innych planach ułamkowych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Zidentyfikowane wartości współczynników modelu nie są tymi, które eksperymentator chciał wyznaczyć, czyli są estymatorami obciążonymi. Zawsze trzeba mieć świadomość, że przy ułamkowych planach eksperymentu uzyskane wartości współczynników modelu są zaburzone pominiętymi współczynnikami regresji.

Chociaż przedstawiona metoda wyznaczania zaburzenia jest dosyć elegancka, to jednak wymaga uwagi przy wyznaczaniu potrzebnych iloczynów macierzy. Znacznie szybciej można ocenić zależności między współczynnikami na podstawie tabeli z macierzą planowania.

Tabela 14.4. Macierz planowania i plan ułamkowego eksperymentu czynnikowego typu 2^{3-1}

Nr	t_0	t_1	t_2	t_3	t_1t_2	t_1t_3	t_2t_3	$t_1t_2t_3$	Kod
1	+1	+1	−1	−1	−1	−1	+1	+1	a
2	+1	−1	+1	−1	−1	+1	−1	+1	b
3	+1	−1	−1	+1	+1	−1	−1	+1	c
4	+1	+1	+1	+1	+1	+1	+1	+1	abc

W tabeli 14.4 przedstawiono plan eksperymentu połówkowego, który został wyznaczony wcześniej. Porównanie kolumn tego planu pozwala stwierdzić, że niektóre kolumny są identyczne. I tak, kolumna t_0 jest identyczna z kolumną $t_1t_2t_3$, kolumna t_1 z kolumną t_2t_3, następnie kolumna t_2 i kolumna t_1t_3, wreszcie t_3 oraz t_1t_2. Przy wyznaczaniu współczynników regresji kolumny liniowo zależne od siebie wpływają na wyznaczane na ich podstawie współczynniki. Widać zatem, że zaburzenia współczynników regresji zapisane w równaniu (14.15) odpowiadają liniowo zależnym kolumnom w macierzy eksperymentu.

Plany eksperymentu można też opisywać równaniami. Na przykład plan przedstawiony w tabeli 14.4 najlepiej charakteryzuje fakt, że wszystkie wiersze planu spełniają równość

$$t_1t_2t_3 = +1$$

co nazywane jest równością charakterystyczną planu lub też kontrastem określającym. Łatwo można sprawdzić, że plan typu 2^{3-1} zakodowany jako [(1), ab, ac, bc] spełnia równość

$$t_1t_2t_3 = -1$$

Wykorzystując równości charakterystyczne, można także określać zdolność rozdzielczą planu ułamkowego.

Na koniec omawiania planów dwupoziomowych jeszcze raz zostanie powtórzona ważna informacja zawarta w poniższej ramce.

Za pomocą planów czynnikowych na dwóch poziomach nie da się identyfikować modeli zawierających człony kwadratowe.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Jeżeli model zawiera człony kwadratowe typu $k_{ii}t_i^2$, to w macierzy planowania kolumny im odpowiadające będą zawierały same wartości +1, co oznacza, że będą identyczne z kolumną t_0 opisującą wyraz wolny funkcji regresji. Zatem nie da się oddzielić wpływu członów kwadratowych od wpływu składowej stałej.

14.4. Plany trójpoziomowe

Stosunkowo rzadko stosowane plany trójpoziomowe pozwalają na wyznaczenie powierzchni odpowiedzi drugiego stopnia. W planowaniu czynnikowym na trzech poziomach eksperyment, w którym każda z S zmiennych wejściowych przyjmuje wartości na trzech poziomach, wymaga wykonania $N = 3^S$ doświadczeń. W praktyce oznacza to, że wzrost liczby zmiennych powoduje gwałtowny, wykładniczy wzrost liczby doświadczeń. W związku z tym plany czynnikowe na trzech poziomach rozpatrywane są jedynie dla niewielkiej liczby zmiennych. Dla dwóch zmiennych trzeba $3^2 = 9$ doświadczeń, dla trzech $3^3 = 27$ doświadczeń, dla czterech wejść to już $3^4 = 81$ doświadczeń. Kolumny macierzy eksperymentu w planach trójpoziomowych nie są ortogonalne względem siebie. Fakt ten stanowi pewne utrudnienie przy rachunkach wykonywanych na papierze i ma konsekwencje w postaci niemożności niezależnego wyznaczania wariancji współczynników regresji.

W odróżnieniu od planów dwupoziomowych, plany trójpoziomowe służą przede wszystkim do identyfikacji modeli matematycznych ze składnikami kwadratowymi, czyli w postaci

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 \tag{14.16}$$

Eksperyment przeprowadzany jest na trzech poziomach, czyli każda ze zmiennych wejściowych przyjmuje trzy różne wartości, maksymalną x_i^{max}, minimalną x_i^{min} lub w środku przedziału zmienności x_i^0. Po wprowadzeniu przyrostu Δx_i wartości zmiennych wejściowych można przedstawić jako

$$\Delta x_i = \frac{x_i^{max} - x_j^{min}}{2} \tag{14.17}$$

$$x_i^0 - \Delta x_i, \quad x_i^0, \quad x_i^0 + \Delta x_i \quad \text{dla } i = 1, 2, \ldots, S \tag{14.18}$$

Podobnie jak przy planach dwupoziomowych można dokonać standaryzacji zmiennych

$$t_i = \frac{x_i - x_i^0}{\Delta x_i}, \quad i = 1, 2, \ldots, S \tag{14.19}$$

Poszczególne zmienne wejściowe standaryzowane przyjmą wartości -1, 0 lub $+1$. Po standaryzacji poszukiwany model matematyczny ma postać

$$\hat{y} = a_0 + a_1t_1 + a_2t_2 + a_{11}t_1^2 + a_{22}t_2^2 + a_{12}t_1t_2 \tag{14.20}$$

Plan i macierz eksperymentu na trzech poziomach dla dwóch zmiennych wejściowych przedstawia tabela 14.5. Jest to plan eksperymentu typu 3^2 i został zilustrowany graficznie na rysunku 14.6.

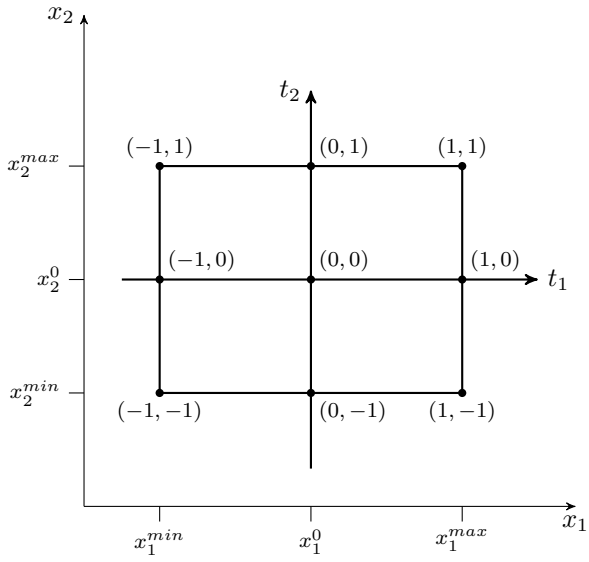
Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 14.5. Macierz i plan eksperymentu czynnikowego typu 3^2

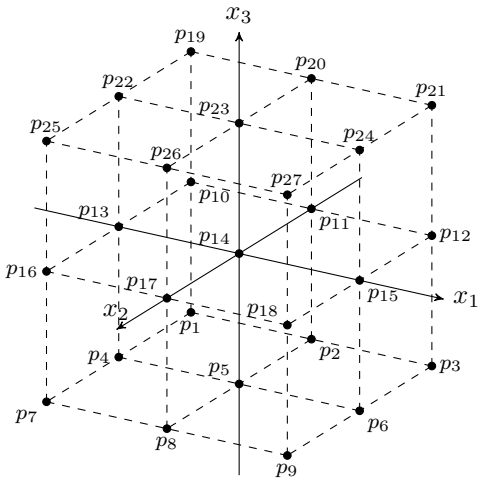
Nr	t_0	t_1	t_2	t_1^2	t_2^2	t_1t_2
1	1	−1	−1	1	1	1
2	1	0	−1	0	1	0
3	1	1	−1	1	1	−1
4	1	−1	0	1	0	0
5	1	0	0	0	0	0
6	1	1	0	1	0	0
7	1	−1	1	1	1	−1
8	1	0	1	0	1	0
9	1	1	1	1	1	1

Jako że obszar badań we współrzędnych naturalnych jest prostokątem, plan eksperymentu zakłada wykonanie doświadczeń w czterech wierzchołkach tego obszaru, następnie w dwóch parach symetrycznie rozłożonych punktów leżących na osiach układu współrzędnych, a także jednego doświadczenia w środku układu współrzędnych. W sytuacji gdy liczba zmiennych wejściowych wyniesie trzy, liczba doświadczeń wzrośnie do 27. Obszar badań we współrzędnych standaryzowanych będzie wtedy sześcianem, w którego środku znajduje się początek układu współrzędnych. Osiem punktów planu: p_1, p_3, p_7, p_9, p_{19}, p_{21}, p_{25} oraz p_{27} znajdować się będzie w wierzchołkach sześcianu. Środki poszczególnych krawędzi wyznaczać będą kolejne 12 doświadczeń, czyli: p_2, p_4, p_6, p_8, p_{10}, p_{12}, p_{16}, p_{18}, p_{20}, p_{22}, p_{24} oraz p_{26}. Następne 6 doświadczeń: p_5, p_{11}, p_{13}, p_{15}, p_{17} i p_{23} znajdować się będzie w środkach każdego z boków sześcianu, a ostatnie doświadczenie p_{14} umieszczone jest w środku układu współrzędnych, jak to pokazano na rysunku 14.7. Jak już wcześniej wspomniano, planowanie czynnikowe na trzech poziomach ma małe zastosowanie praktyczne. Jest to związane z wykładniczym wzrostem liczby doświadczeń w miarę wzrostu liczby zmiennych wejściowych. W związku z tym plany tego rodzaju stosuje się wyłącznie w przypadkach, gdy liczba zmiennych S wynosi 2 bądź 3. W przypadku większej liczby zmiennych zalecane jest stosowanie planów kompozycyjnych, w szczególności planowania pięciopoziomowego, którego podstawą jest planowanie dwupoziomowe uzupełnione o tak zwane doświadczenie gwiezdne, w którym zmienia się kolejno zmienne na dwóch poziomach, oraz doświadczenia centralne w środku planu eksperymentu. Wyznaczenie współczynników funkcji regresji przeprowadza się w sposób identyczny jak przy planowaniu dwupoziomowym.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 14.6. Interpretacja graficzna planu eksperymentu czynnikowego na trzech poziomach typu 3^2



Rysunek 14.7. Graficzna interpretacja planu eksperymentu czynnikowego na trzech poziomach typu 3^3

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

14.5. Plany Placketta–Burmana

Chociaż należałoby mówić raczej o metodzie Placketta–Burmana, zwyczajowo używa się określenia „plany". Plany Placketta–Burmana zostały opracowane w celu minimalizacji wariancji wyznaczonych współczynników regresji przy możliwie małej liczbie doświadczeń. Można je budować dla różnej liczby poziomów wartości wejściowych. Najpopularniejsze są plany dwupoziomowe.

Robin L. Plackett i J. P. Burman zajęli się opracowaniem planu eksperymentu, który pozwoliłby dokonać pewnej optymalizacji i ocenić wpływ zmian wprowadzanych w poszczególnych elementach pewnego zespołu złożonego z N elementów. Przy czym każdy z tych elementów może być w dwóch stanach x_i^0 zwanym stanem podstawowym oraz x_i', który jest stanem zmodyfikowanym. Ocena wpływu poszczególnych elementów na wartość zespołu przeprowadzana jest w postaci ilorazu

$$m_i = \frac{\sum y(x_1', x_2, \ldots, x_N) - \sum y(x_1^0, x_2^0, \ldots, x_N^0)}{2^N}$$

gdzie $y(x_1, x_2, \ldots, x_N)$ jest mierzalną wartością zespołu, zależną od sposobu złożenia, $y(x_1^0, x_2^0, \ldots, x_N^0)$ jest pewnym stanem podstawowym zespołu podlegającego modyfikacjom związanym ze zmianami poszczególnych elementów.

Najważniejszą cechą tych planów jest pomijanie interakcji między czynnikami. Plany dwupoziomowe Placketta–Burmana budowane są tak, by dla każdej pary czynników każda kombinacja poziomów występowała tyle samo razy w całym eksperymencie. Pełne plany czynnikowe spełniają to kryterium, lecz plany Placketta–Burmana wymagają mniejszej liczby doświadczeń. Dla liczby doświadczeń $N = 2^S$, plany Placketta–Burmana pokrywają się z planami czynnikowymi na dwóch poziomach całkowitymi czy ułamkowymi. Plany Placketta–Burmana można budować dla N doświadczeń, przy czym N musi być wielokrotnością liczby cztery. Dlatego też najczęściej się je stosuje, gdy liczba doświadczeń jest wielokrotnością 4, ale nie jest całkowitą potęgą liczby 2. Maksymalna liczba czynników, których parametry regresji się wyznacza, jest o jeden mniejsza niż liczba doświadczeń. Jeżeli liczba czynników branych pod uwagę w analizie jest mniejsza niż $N - 1$, to nadmiarowe kolumny się odrzuca, przy czym można ich użyć do określenia jakości dopasowania funkcji regresji do rzeczywistej odpowiedzi układu. Czyli dla 21 parametrów bierze się plan dla $N = 24$ i odrzuca dwie ostatnie kolumny. Z kolei dla 24 parametrów należy wziąć plan z 28 doświadczeniami i odrzucić trzy kolumny.

Cenną cechą tych planów jest fakt, że kolumny macierzy eksperymentu są ortogonalne względem siebie. Dla dwóch poziomów wartości plany Placketta–Burmana są macierzami Hadamarda. Tak więc są one kwadratowe, zaś liczba doświadczeń N jest o jeden większa od liczby wyznaczanych parametrów.

14.6. Plany wielopoziomowe

Powstały jako odpowiedź na potrzebę budowy modeli matematycznych ze składnikami kwadratowymi dla układów z liczbą wejść większą niż trzy. Sposób tworzenia

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

planów wielopoziomowych odbiega od tego, co można było obserwować przy planach typu 2^S oraz 3^S. Aby ograniczyć liczbę wykonywanych doświadczeń, wykorzystuje się plany:

— kompozycyjne,
— ortogonalne,
— rotatabilne.

Nazwa kompozycyjny dotyczy planu, w którym promień gwiezdny jest przyjmowany w zasadzie w dowolny sposób. Plany ortogonalne mają diagonalną macierz informacyjną $\mathbb{X}^T\mathbb{X}$. Natomiast rotatabilność czy, jak niektórzy autorzy piszą, rotalność planu oznacza niezależność jakości wyznaczenia współczynników regresji od obrotu przyjętego układu współrzędnych, co oznacza niezmienniczość macierzy kowariancji $(\mathbb{X}^T\mathbb{X})^{-1}$ względem obrotu układu współrzędnych. Plan ortogonalny ma jedno doświadczenie centralne i odpowiednio dobrany promień gwiezdny. Plan rotatabilny ma odpowiedni promień gwiezdny i odpowiednią liczbę doświadczeń centralnych.

Zatem powyższy podział ma znaczenie historyczne. Ortogonalność planu miała istotne znaczenie w czasach, gdy obliczenia wykonywane były np. tylko za pomocą suwaka logarytmicznego. Postęp techniki obliczeniowej spowodował spadek znaczenia planów ortogonalnych. Poniżej zostaną pokrótce omówione wybrane zagadnienia związane z posługiwaniem się tymi planami.

14.6.1. Plany kompozycyjne

Plany kompozycyjne powstają z planów dwupoziomowych całkowitych typu 2^S lub ułamkowych typu 2^{S-P}, uzupełnionych o tak zwane doświadczenia centralne oraz doświadczenie gwiezdne. Doświadczenie gwiezdne polega na zmianie poszczególnych zmiennych na dwóch poziomach $\pm a$, przy czym pozostałe zmienne są na poziomie zero. Oznacza to, że punkty gwiezdne leżą na osiach współrzędnych, ułożone symetrycznie po dwa na każdej z osi, w odległości $\pm a$ od środka planu. Odległość a jest nazywana promieniem gwiezdnym.

Zazwyczaj plany kompozycyjne buduje się z użyciem zmiennych standaryzowanych. Całkowita liczba doświadczeń planu kompozycyjnego wynika z liczby doświadczeń planu czynnikowego całkowitego i wynosi

$$N = 2^S + 2S + N_0 \tag{14.21}$$

gdzie S jest liczbą czynników, N_0 oznacza liczbę doświadczeń centralnych. Natomiast w przypadku planów ułamkowych 2^{S-P}

$$N = 2^{S-P} + 2S + N_0 \tag{14.22}$$

gdzie P jest ułamkowością planu. Zatem na całkowitą liczbę doświadczeń składają się następujące składniki:

— 2^S lub 2^{S-P} doświadczeń wykonanych zgodnie z planem czynnikowym całkowitym lub ułamkowym na dwóch poziomach,

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

— $2S$ symetrycznie rozłożonych wokół środka planu doświadczeń gwiezdnych typu $(0, \ldots, 0 \pm a, 0, \ldots, 0)$, gdzie wartość $\pm a$ przyjmują kolejne zmienne wejściowe,
— N_0 tak zwanych doświadczeń centralnych typu $(0, 0, \ldots, 0)$ wykonanych w środku planu.

Chyba najczęściej wykorzystywanymi planami wielopoziomowymi są plany centralne kompozycyjne, zwane też planami Boxa–Wilsona. Istnieją trzy rodzaje planów centralnych kompozycyjnych: opisane, wpisane i z węzłami umieszczonymi na ścianach.

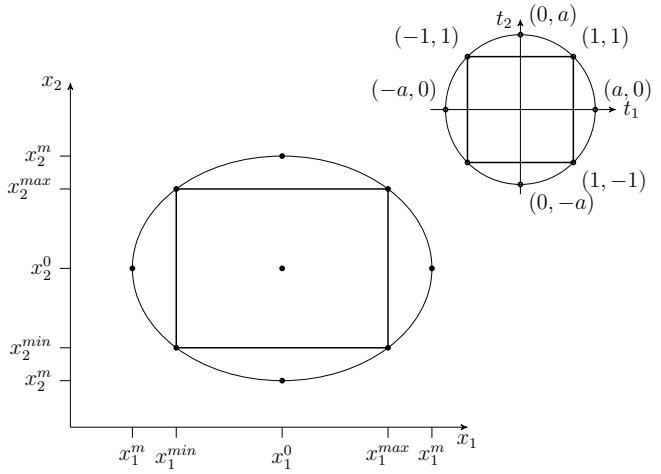
Plany dla dwóch i trzech czynników zamieszczono w tabelach odpowiednio 14.6 oraz 14.7. W obu planach wyraźnie widać widać powyższy podział. W planie dla dwóch czynników (tab. 14.6) 4 wiersze to plan całkowity typu 2^2, wiersze od 5 do 8 to doświadczenia gwiezdne, zaś wiersz 9 to doświadczenie centralne. Jeżeli długość promienia gwiezdnego a wynosi $\sqrt{2}$, to ilustracją planu z tabeli 14.6 jest rysunek 14.8. Przedstawia on plan centralny kompozycyjny we współrzędnych naturalnych x_1, x_2, gdzie punkty planu należą do pewnej elipsy, oraz we współrzędnych standaryzowanych, gdzie układają się w okrąg. W prawym górnym rogu rysunku 14.8 widać zawartość macierzy planu w postaci punktów. Wyraźnie widać kwadrat wynikający z planu całkowitego 2^2, doświadczenia gwiezdne oraz centralne. Plan ujęty w tabeli 14.6 oraz na rysunku 14.8 nosi nazwę centralnego kompozycyjnego opisanego (CCC) (*Central Composite Circumscribed*).

Tabela 14.6. Plan eksperymentu kompozycyjnego dla dwóch czynników

Nr	t_0	t_1	t_2	t_1^2	t_2^2	$t_1 t_2$	y
1	$+1$	-1	-1	$+1$	$+1$	$+1$	y_1
2	$+1$	$+1$	-1	$+1$	$+1$	-1	y_2
3	$+1$	-1	$+1$	$+1$	$+1$	-1	y_3
4	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	y_4
5	$+1$	$-a$	0	a^2	0	0	y_5
6	$+1$	$+a$	0	a^2	0	0	y_6
7	$+1$	0	$-a$	0	a^2	0	y_7
8	$+1$	0	$+a$	0	a^2	0	y_8
9	$+1$	0	0	0	0	0	y_9

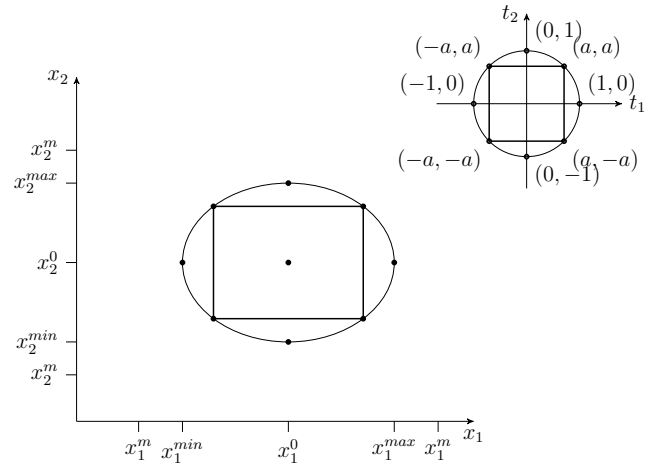
Sama nazwa jest oczywista i jasno wynika z rysunku. Warto zwrócić uwagę na różnice między rysunkami 14.8 oraz 14.9.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356



Rysunek 14.8. Interpretacja graficzna centralnego kompozycyjnego opisanego (CCC) planu eksperymentu na pięciu poziomach we współrzędnych naturalnych i standaryzowanych dla dwóch zmiennych

Na podstawie tabeli 14.7 można prześledzić zmiany wartości czynników i członów kwadratowych oraz członów opisującymi interakcje w planie centralnym kompozycyjnym opisanym dla trzech czynników. Dla trzech czynników trzeba przynajmniej piętnastu doświadczeń, by dokonać identyfikacji współczynników powierzchni odpowiedzi.



Rysunek 14.9. Interpretacja graficzna centralnego kompozycyjnego wpisanego (CCI) planu eksperymentu na pięciu poziomach we współrzędnych naturalnych i standaryzowanych dla dwóch zmiennych

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 14.7. Plan eksperymentu kompozycyjnego (CCC) dla trzech czynników

Nr	t_0	t_1	t_2	t_3	t_1^2	t_2^2	t_3^2	t_1t_2	t_1t_3	t_2t_3	$t_1t_2t_3$	y
1	+1	−1	−1	−1	+1	+1	+1	+1	+1	+1	−1	y_1
2	+1	+1	−1	−1	+1	+1	+1	−1	−1	+1	+1	y_2
3	+1	−1	+1	−1	+1	+1	+1	−1	+1	−1	+1	y_3
4	+1	+1	+1	−1	+1	+1	+1	+1	−1	−1	−1	y_4
5	+1	−1	−1	+1	+1	+1	+1	+1	−1	−1	+1	y_5
6	+1	+1	−1	+1	+1	+1	+1	−1	+1	−1	−1	y_6
7	+1	−1	+1	+1	+1	+1	+1	−1	−1	+1	−1	y_7
8	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	y_8
9	+1	$-a$	0	0	a^2	0	0	0	0	0	0	y_9
10	+1	$+a$	0	0	a^2	0	0	0	0	0	0	y_{10}
11	+1	0	$-a$	0	0	a^2	0	0	0	0	0	y_{11}
12	+1	0	$+a$	0	0	a^2	0	0	0	0	0	y_{12}
13	+1	0	0	$-a$	0	0	a^2	0	0	0	0	y_{13}
14	+1	0	0	$+a$	0	0	a^2	0	0	0	0	y_{14}
15	+1	0	0	0	0	0	0	0	0	0	0	y_{15}

Rysunek 14.9 także przedstawia ilustrację centralnego planu kompozycyjnego. Różnica polega na wybraniu wartości maksymalnych. W planie centralnym kompozycyjnym opisanym (CCC) wartości zmiennych w doświadczeniach gwiezdnych wychodzą poza przyjęty zakres, odwzorowywany przez zmienne standaryzowane $t_i \in \langle -1, 1\rangle$.

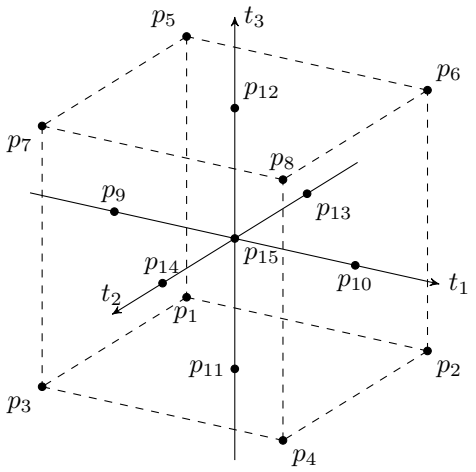
Z tego powodu w niektórych zastosowaniach plan CCC nie może być stosowany, ponieważ nie ma fizycznej możliwości nadania tak dużych wartości wejściom układu. W takim przypadku można zastosować plan centralny kompozycyjny wpisany (CCI) (*Central Composite Inscribed*) przedstawiony schematycznie na rysunku 14.9. Zasadniczą zaletą planu CCI jest respektowanie ograniczenia dotyczącego maksymalnej i minimalnej wartości czynników. Ostatnim rodzajem planu centralnego kompozycyjnego jest (CCF) (*Central Composite Face Centered*) – plan z wartościami gwiezdnymi na ścianach hiperkostki stanowiącej geometryczną reprezentację planu. Taką reprezen-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

tację dla trzech czynników zamieszczono w tabeli 14.8. Po lewej stronie znajdują się standaryzowane wartości przyjmowane przez poszczególne czynniki, zaś po prawej jest rysunek przedstawiający położenie punktów opisanych kolejnymi wierszami planu. Wybrane cechy planów centralnych kompozycyjnych zestawiono w tabeli 14.9.

Tabela 14.8. Plan centralny kompozycyjny z doświadczeniami na ścianach (CCF) dla trzech czynników

Nr	t_0	t_1	t_2	t_3	y	Schemat planu
1	1	−1	−1	−1	y_1	
2	1	1	−1	−1	y_2	
3	1	−1	1	−1	y_3	
4	1	1	1	−1	y_4	
5	1	−1	−1	1	y_5	
6	1	1	−1	1	y_6	
7	1	−1	1	1	y_7	
8	1	1	1	1	y_8	
9	1	−1	0	0	y_9	
10	1	1	0	0	y_{10}	
11	1	0	−1	0	y_{11}	
12	1	0	1	0	y_{12}	
13	1	0	0	−1	y_{13}	
14	1	0	0	1	y_{14}	
15	1	0	0	0	y_{15}	



Jak już wcześniej wspomniano, plany kompozycyjne mogą być ortogonalne lub rotatabilne. Ortogonalność planu oznacza, że macierz informacyjna $\mathbb{X}^T\mathbb{X}$ jest diagonalna, co wymaga zerowania się iloczynów skalarnych poszczególnych kolumn macierzy eksperymentu. Ortogonalność planu osiąga się, dobierając odpowiednio długość ramienia gwiezdnego w planach kompozycyjnych.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

I tak, długość ramienia gwiezdnego dla planów ortogonalnych zbudowanych z wykorzystaniem całkowitego planu czynnikowego dwupoziomowego powinna wynosić

$$a = \sqrt{\frac{1}{2}\left[\sqrt{2^S\left(2^S + 2S + N_0\right)} - 2^S\right]}$$

Natomiast dla planów ortogonalnych kompozycyjnych budowanych na podstawie planów ułamkowych długość promienia gwiezdnego wynosi

$$a = \sqrt{\frac{1}{2}\left[\sqrt{2^{S-P}\left(2^{S-P} + 2S + N_0\right)} - 2^{S-P}\right]}$$

Od czasu, gdy komputery stały się podstawowym narzędziem pracy inżyniera, ortogonalność planu, która prowadzi do znacznego uproszczenia obliczeń, nie jest aż tak istotna. Znacznie ważniejsze staje się uzyskiwanie oszacowań współczynników funkcji regresji, które nie zależy od obrotu układu współrzędnych planu.

14.6.2. Planowanie rotatabilne

Główną wadą opisywanego kompozycyjnego planu eksperymentu jest różna wariancja wartości y, która jest przewidywana na podstawie równania funkcji regresji (14.16).

Plany kompozycyjne można opracowywać w taki sposób, by zapewniały rotatabilność (rotalność), co oznacza niezależność jakości oszacowania współczynników regresji od obrotu układu współrzędnych. Aby osiągnąć taką właściwość, plan musi mieć odpowiedni promień gwiezdny oraz odpowiednią liczbę doświadczeń centralnych. Przy planach budowanych na podstawie całkowitych dwupoziomowych planów czynnikowych promień gwiezdny dla planu rotatabilnego powinien wynosić

$$a = \sqrt[4]{2^S} \tag{14.23}$$

W przypadku, gdy $S \geqslant 5$, efektywniejsze jest wykorzystanie ułamkowego planu dwupoziomowego typu 2^{S-P}. Wtedy warunek rotatabilności jest spełniony, gdy promień gwiezdny ma wartość

$$a = \sqrt[4]{2^{S-P}} \tag{14.24}$$

Natomiast liczba doświadczeń centralnych dla planu rotatabilnego powinna spełniać warunek

$$\frac{(S+2)(2^S + 2S + N_0)}{S(2^S + 4\sqrt{2^S} + 4)} > 1$$

W niektórych wypadkach powyższy warunek nie jest wystarczający. Na przykład, czasem można wymagać, by wariancja odpowiedzi modelu w punkcie centralnym planu była identyczna z wariancją w punktach leżących na powierzchni kuli o promieniu jednostkowym; w takiej sytuacji rośnie liczba doświadczeń centralnych i osiąga wartości pokazane poniżej. Zestawienie zawiera długości promienia gwiezdnego a, liczby doświadczeń centralnych N_0 oraz całkowite liczby doświadczeń N w zależności od liczby czynników (zmiennych wejściowych) S w eksperymencie.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

S	2	3	4	5	6
a	1,414	1,682	2	2,378	2,828
N_0	5	6	7	10	15
N	13	20	31	52	91

Widać silną zależność między liczbą czynników a liczbą doświadczeń centralnych wymaganą przy planach rotatabilnych. Między innymi z powodu tej zależności tego rodzaju plany nie znajdują dziś szerokiego zastosowania do eksperymentów numerycznych, czyli takich, które polegają na badaniach symulacyjnych obiektów.

Warto je stosować w przypadku, gdy zastosowane modele obliczeniowe nie będą deterministyczne i pozwolą na uzyskiwanie losowych wyników.

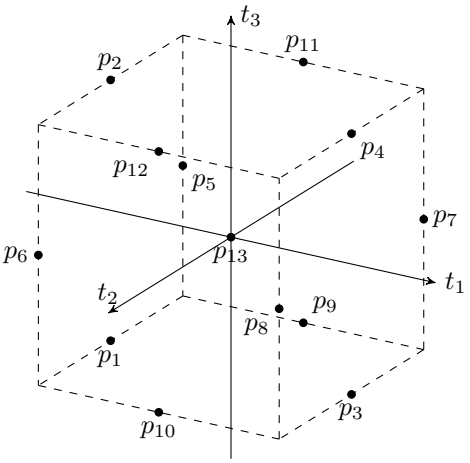
Tabela 14.9. Właściwości planów wielopoziomowych

Rodzaj planu	Cechy	Liczba poziomów
CCC	Pozwala na uzyskanie wysokiej jakości dopasowania powierzchni odpowiedzi w całym badanym obszarze, niestety wymaga wartości parametrów spoza zakresu przyjmowanego przez czynnikową część planu.	5
CCI	Wykorzystuje tylko punkty wewnątrz obszaru przyjętego początkowo dla części czynnikowej planu. Zapewnia mniejszą dokładność estymacji parametrów w całym badanym obszarze niż CCC.	5
CCF	Stosunkowo wysoka jakość estymacji w całym badanym obszarze, nie wymaga doświadczeń poza zakresem badań, daje kiepskie oszacowania czysto kwadratowych zależności.	3
Box–Behnken	Wymaga mniej doświadczeń niż plany centralne kompozycyjne dla przypadków z liczbą czynników mniejszą lub równą 4. Jest planem rotatabilnym, ale w niektórych obszarach daje kiepskie oszacowania współczynników regresji podobnie jak CCI, brak eksperymentów w narożach planu może być użyteczny, gdy badany obiekt źle znosi równoczesne przyjmowanie wartości ekstremalnych przez kilka czynników.	3

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Tabela 14.10. Plan Boxa–Behenkena dla trzech czynników

Nr	t_0	t_1	t_2	t_3	y	Schemat planu
1	+1	−1	−1	0	y_1	
2	+1	−1	+1	0	y_2	
3	+1	+1	−1	0	y_3	
4	+1	+1	+1	0	y_4	
5	+1	−1	0	−1	y_5	
6	+1	−1	0	+1	y_6	
7	+1	+1	0	−1	y_7	
8	+1	+1	0	+1	y_8	
9	+1	0	−1	−1	y_9	
10	+1	0	−1	+1	y_{10}	
11	+1	0	+1	−1	y_{11}	
12	+1	0	+1	+1	y_{12}	
13	+1	0	0	0	y_{13}	
14	+1	0	0	0	y_{14}	
15	+1	0	0	0	y_{15}	



14.6.3. Plany Boxa–Behnkena

Plany Boxa–Behnkena są planami trójpoziomowymi. Podobnie jak plany centralne kompozycyjne, plany te można wykorzystywać do wyznaczania funkcji regresji ze składnikami kwadratowymi. Są to plany rotatabilne lub prawie rotatabilne, zaś dla liczby czynników mniejszej niż cztery wymagają mniej doświadczeń niż plany centralne kompozycyjne. Przykładowy plan dla trzech czynników przedstawiono w tabeli 14.10. To, że w doświadczeniach czynniki nie przyjmują jednocześnie ekstremalnych wartości, bywa w niektórych przypadkach znaczną zaletą. Natomiast wadą tych planów jest to, że nie da się precyzyjnie oszacować ekstremów powierzchni odpowiedzi.

Podobnie jak plany centralne kompozycyjne wpisane nie dają pełnej informacji na temat zachowania badanego obiektu na skraju analizowanego obszaru. Tabela 14.9 zawiera porównanie tych planów z planami centralnymi kompozycyjnymi.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

14.7. Planowanie losowe

Budowanie modelu zastępczego w postaci funkcji regresji jest obarczone dużą niepewnością. Wiąże się z wyborem zmiennych decyzyjnych, przy czym nie zawsze można z całą pewnością określić, które zmienne są najważniejsze. Oprócz tego obiekty rzeczywiste mogą być wrażliwe na kolejność zmiany stanów wejść. Mogą też mieć miejsce zakłócenia skorelowane z niektórymi wejściami. Aby zminimalizować wpływ takich zjawisk na otrzymany model, można wykorzystać randomizację. Oznacza ona że doświadczenia w eksperymencie będą przeprowadzane w losowej kolejności.

Plan randomizowany bardzo dobrze sprawdza się przy weryfikacji istotności wpływu wielkości wejściowych na wielkość wyjściową, ponieważ plan ten wprowadza element losowości do zbioru punktów pomiarowych.

Losowość do planów eksperymentu można wprowadzać na kilka sposobów. Jednym ze sposobów wprowadzania losowości do planów jest metoda Monte Carlo polegająca na losowym wyborze wartości poszczególnych parametrów z badanego obszaru.

Z obliczeniowego punktu widzenia wydajniejsza jest metoda kwadratów łacińskich (*Latin Hypercube sampling*). Ogólna idea metody jest następująca: cały analizowany obszar przestrzeni parametrów wejściowych przedstawia się w postaci wielowymiarowej kostki. Dla dwóch czynników ta kostka przyjmuje postać kwadratu. Następnie dzieli się uzyskaną kostkę na przedziały wartości o jednakowym prawdopodobieństwie występowania każdej ze zmiennych. W dwuwymiarowym przypadku prowadzi to do uzyskania jednakowej liczby kolumn i wierszy. Następnie dla każdego wiersza losuje się pozycję próbki w ten sposób, by w każdej kolumnie znajdowała się dokładnie jedna próbka. Dla wielowymiarowego przypadku postępuje się podobnie.

Metoda kwadratów łacińskich wymaga mniejszej liczby próbek niż metoda Monte Carlo, do uzyskania podobnych niepewności wyników, zatem jest ekonomiczniejsza z obliczeniowego punktu widzenia.

14.8. Wspomaganie realizacji eksperymentów komputerowych

Wspomaganie eksperymentów przeprowadzanych metodami numerycznymi za pomocą komputerów najczęściej jest realizowane przez wyspecjalizowane moduły pakietów obliczeniowych. Takie moduły mają, między innymi: ANSYS, Comsol, Abaqus i wiele innych. W zasadzie w każdym z powszechnie dostępnych pakietów do rozwiązywania problemów inżynierskich metodami numerycznymi znajdują się tego typu narzędzia. Ich zasadniczą zaletą jest dobra integracja z resztą pakietu, zaś wadą jest to, że każda z nich stanowi odrębną implementację metod optymalizacji i planowania eksperymentu. Użytkownik jest w pewien sposób skazany na ofertę producenta i zazwyczaj tylko w ograniczonym stopniu może modyfikować ustawienia i używane algorytmy.

W przypadku niektórych programów obliczeniowych użytkownik może sam napisać odpowiedni program do realizacji eksperymentu numerycznego. Można tego dokonać na przykład z poziomu MATLAB-a czy innego środowiska obliczeniowego. Można

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

też wykorzystać w zasadzie dowolny język programowania od C, przez FORTRAN aż po Python.

Oprócz tych specjalizowanych modułów czy procedur pisanych przez badaczy istnieją specjalizowane programy służące tylko do wspomagania eksperymentów i optymalizacji. Jednym z nich jest rozwijany przez Sandia National Laboratories pakiet DAKOTA. Na następnych stronach zostanie przybliżony sposób pracy z tym pakietem, jego możliwości i sposób wykorzystania do realizacji eksperymentów numerycznych z pomocą MATLAB-a czy GNU Octave. Można go też użyć do przeprowadzania eksperymentów komputerowych z wykorzystaniem dowolnego oprogramowania obliczeniowego. DAKOTA wykorzystuje podejście określane mianem „czarnej skrzynki", co oznacza, że interakcja z modelem odbywa się przez zmiany parametrów wejściowych badanego obiektu np. zapisanych w pliku. Nie ma wymagań dotyczących implementacji modelu czy sposobu komunikowania się, dzięki temu jest to bardzo elastyczny pakiet.

Jeżeli tylko można napisać program czy skrypt, który pozwoli na sterowanie parametrami modelu, to można wykorzystać pakiet DAKOTA do wspomagania lub realizacji danego eksperymentu numerycznego.

14.9. Wykorzystanie pakietu DAKOTA

DAKOTA jest oprogramowaniem, które dostarcza użytkownikowi różnorodne metody iteracyjne i algorytmy obliczeniowe. Pozwala na elastyczne włączanie ich w proces symulacji z wykorzystaniem MATLAB-a, GNU Octave, Scilaba, Pythona itp. Główne klasy algorytmów zostały przedstawione poniżej.

— **Analiza parametrów** (*Parameter Studies*) pozwala określić wpływ zmian wartości wejściowych na wartości wyjściowe w parametrycznych modelach symulacyjnych. Może być pomocna w ocenie cech modeli symulacyjnych, takich jak gładkość, multimodalność czy nieliniowość, które z kolei mają wpływ na wybór algorytmów i mechanizmów kontroli w ewentualnych dalszych pracach nad optymalizacją i analizą niepewności. Typowe przykłady obejmują zmiany jednego czynnika na raz lub jednoczesne zmiany wartości parametrów.

— **Planowanie eksperymentu** (*Design of Exepriments*) jest używane do określania przestrzeni parametrów wejściowych danego problemu inżynierskiego, na przykład w celu przeprowadzenia analizy wrażliwości globalnej. Planowanie eksperymentów numerycznych pozwala na uzyskanie informacji podobnych do tych z analizy parametrów, jednak zasadniczym celem tych algorytmów jest jak najlepsze przebadanie całej przestrzeni parametrów wejściowych. Najczęściej wykorzystuje się kwadraty łacińskie i plany Boxa–Behenkena.

— **Analiza niepewności** (*Uncertainty Quantification*) służy do wyznaczania statystycznych charakterystyk opisujących odpowiedzi układu na podstawie symulacji przeprowadzonych według określonych rozkładów prawdopodobieństwa parametrów wejściowych.

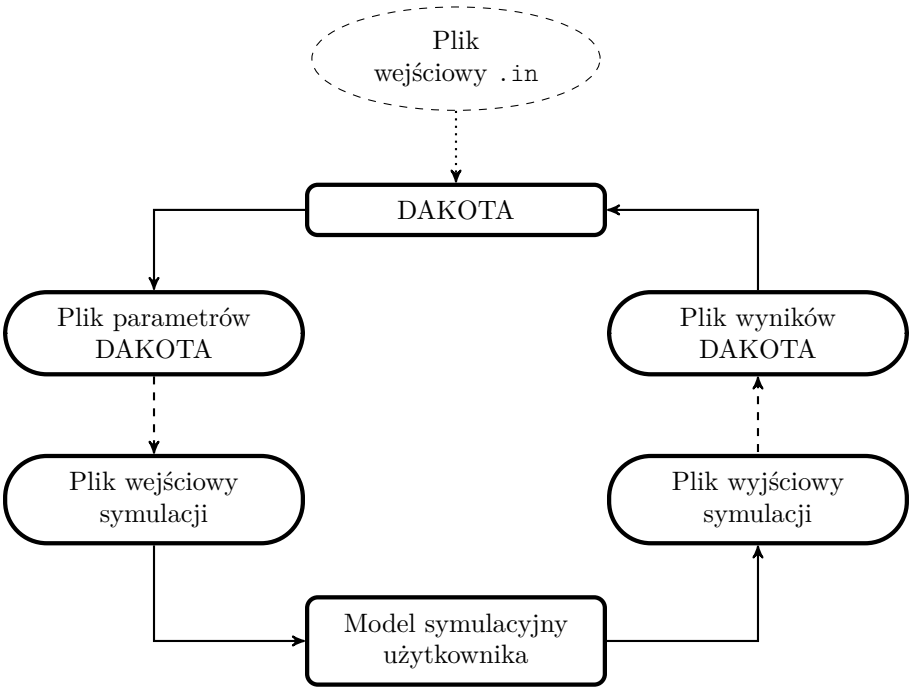
— **Optymalizacja** (*Optimization*) obejmuje algorytmy dążące do minimalizacji funkcji celu, która często obejmuje koszty oraz wydajność układu, zgodnie z przewi-

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

dywaniami modelu symulacyjnego. Optymalizacja uwzględnia ograniczenia dotyczących zmiennych wejściowych lub odpowiedzi układu.

— **Kalibracja** (*Calibration*) służy do maksymalizacji zgodności pomiędzy wartościami wyjściowymi z obiektu numerycznego a eksperymentalnymi danymi dotyczącymi obiektu rzeczywistego. Najczęściej stosuje się je w celu rozwiązywania problemów odwrotnych to znaczy estymacja parametrów modelu. DAKOTA używa nieliniowej metody najmniejszych kwadratów oraz metod bayesowskich.

Zastosowane w pakiecie DAKOTA interfejsy pozwalają na wykorzystanie dowolnego języka programowania do budowy modelu symulacyjnego. Współpraca między DAKOTĄ a modelem symulacyjnym przebiega zgodnie z rysunkiem 14.10.



Rysunek 14.10. Idea wymiany informacji między pakietem DAKOTA a modelem

Główną zaletą DAKOTY jest stosunkowo prosty sposób łączenia jej z modelem opracowanym przez użytkownika. Skutkuje to dosyć luźnym sprzężeniem między dwoma środowiskami obliczeniowymi. Jak widać na rysunku 14.10, sprzężenie odbywa się za pomocą niewielkich plików. DAKOTA sterowana za pomocą poleceń z pliku z rozszerzeniem `.in` zapisuje do pliku parametrów wartości poszczególnych parametrów. Wyniki wygenerowane przez kod użytkownika są odczytywane z pliku wyników. Linie ciągłe na rysunku 14.10 oznaczają operacje wejścia i wyjścia realizowane przez pakiet DAKOTA lub program obliczeniowy użytkownika. Kreskowe linie oznaczają przekazywanie informacji konieczne do zaimplementowania przez użytkownika. Szczegółowe

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

informacje na temat sposobu tworzenia odpowiednich procedur można poznać, studiując odpowiednie instrukcje obsługi stanowiące cześć pakietu.

14.10. Podsumowanie

W niniejszym rozdziale przedstawiono zarys statystyki. Omówiono ideę planowania eksperymentu oraz podstawowe metody, pozwalające budować plany eksperymentu. Czytelnik, który uważnie prześledził ten rozdział, ma wiedzę na temat planów eksperymentu. Jak przekonać się do jej stosowania w praktyce? W końcowym rozdziale publikacji [34] zamieszczono ciekawy przykład ujawniający korzyści z planowania eksperymentu. Warto zapoznać się z tym przykładem, by uzyskać jaśniejszy obraz korzyści z nabytej wiedzy.

Zamysłem autorów było, by niniejszy rozdział stanowił zachętę do dalszego zgłębiania tematu jednocześnie dając pewne podstawy czytelnikowi, który styka się z tematem planowania eksperymentu po raz pierwszy.

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Bibliografia

[1] Achiezer N. I.: *Teoria aproksymacji*, PWN, Warszawa 1953

[2] Becker A. A.: *The Boundary element method in engineering*, McGraw-Hill Book Company, London 1992

[3] Beer G., Watson J. O.: *Introduction to finite and boundary element methods for engineers*, John Wiley & Sons, Chichester 1994

[4] Bhatti M. A.: *Fundamental finite element analysis and applications with mathematica and matlab computations*, John Wiley & Sons, Hoboken 2005

[5] Björk Å., Dahlquist G.: *Metody numeryczne*, PWN, Warszawa 1987

[6] Box G. E. P., Wilson K. B.: *On the experimental attainement of optimum conditions*, Journal of the Royal Statistical Society, Series B, vol. 13, No. 1 (1951), s. 1–45

[7] Brański A.: *Metody numeryczne rozwiązywania zagadnień brzegowych: klasyfikacja i przegląd*, Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów 2013

[8] Butcher J. C.: *Implicit Runge–Kutta Processes*, American Mathematics Society, Mathematics of Computation, vol. 18, 1964, s. 50–64

[9] Butcher J. C.: *A history of Runge–Kutta methods*, Appfied Numerical Mathematics, vol. 20, 1996, s. 247–260

[10] Cannon R. H. Jr: *Dynamika układów fizycznych*, WNT, Warszawa 1973

[11] Dahlquist G.: *A special stability problem for linear multistep methods*, BIT Numerical Mathematics, vol. 3, 1963, s. 27–43, doi=10.1007/BF01963532

[12] Dryja M., Jankowska J., Jankowski M.: *Przegląd metod i algorytmów numerycznych, Część 2*, WNT, Warszawa 1982

[13] Dudek-Dyduch E., Wąs J., Dutkiewicz L., Grobler-Dębska K., Gudowski B.: *Metody numeryczne. Wybrane zagadnienia*, Wydawnictwa AGH, Kraków 2011

[14] Fortuna S.: *Wentylatory. Podstawy teoretyczne, zagadnienia konstrukcyjno-eksploatacyjne i zastosowanie*, Techwent, Kraków 1999

[15] Fortuna Z., Macukow B., Wasowski J.: *Metody numeryczne*, WNT, Warszawa 2005

[16] Gawęcki A.: *Mechanika materiałów i konstrukcji prętowych*, Wyd. Alma Mater Politechniki Poznańskiej, Poznań 2003

[17] Godziszewski J., Mania R., Pampuch R.: *Zasady planowania doświadczeń i opracowywania wyników pomiaru*, Wydawnictwa AGH, Kraków 1982

[18] Hairer, E., Nørsett, S. P., Wanner, G.: *Solving Ordinary Differential Equations I, 2nd Revised. Ed.: Nonstiff Problems*, Springer-Verlag, New York 1993

[19] Hairer, E., Wanner, G.: *Solving Ordinary Differential Equations II, 2nd Revised. Ed.: Stiff and Differential Algebraic Problems*, Springer-Verlag, New York 1996

[20] Jankowska J., Jankowski M.: *Przegląd metod i algorytmów numerycznych. Część 1*, WNT, Warszawa 1988

[21] Jaworski A.:*Metoda elementów brzegowych. Zagadnienia potencjalne*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2000

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

[22] Kiełbasiński A., Schwetlick H.: *Numeryczna algebra liniowa. Wprowadzenie do obliczeń zautomatyzowanych*, WNT, Warszawa 1992
[23] Kincaid D., Cheney W.: *Analiza numeryczna*, WNT, Warszawa 2005
[24] Kuhn, H. W., Tucker, A. W.: *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley 1951, s. 481–492
[25] Kleiber M. (red.): *Metody komputerowe mechaniki ciał stałych*, PWN, Warszawa 1995
[26] Kleiber M.: *Wprowadzenie do metody elementów skończonych*, PWN, Warszawa-Poznań 1989
[27] Korzyński M.: *Metodyka eksperymentu. Planowanie, realizacja i statystyczne opracowanie wyników eksperymentów technologicznych*, WNT, Warszawa 2013
[28] Krupka J., Morawski R.Z., Opalski L.J.: *Metody numeryczne dla studentów elektroniki i technik informacyjnych*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 1997
[29] Krupowicz A. *Metody numeryczne zagadnień początkowych równań różniczkowych zwyczajnych*, PWN, Warszawa 1986
[30] Legras J.: *Praktyczne metody analizy numerycznej*, WNT, Warszawa 1974
[31] Łodygowski T., Kąkol W.:*Metoda elementów skończonych w wybranych zagadnieniach mechaniki konstrukcji inżynierskich*, Wydawnictwo Politechniki Poznańskiej Alma Mater, Poznań 2003
[32] Majchrzak E.: *Metody elementów brzegowych w przepływie ciepła*, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2001
[33] Majchrzak E., Mochnacki B.: *Metody numeryczne. Podstawy teoretyczne, aspekty praktyczne i algorytmy*, wyd. 4, Wydawnictwo Politechniki Śląskiej, Gliwice 2004
[34] Mańczak K.: *Technika planowania eksperymentu*, WNT, Warszawa 1976
[35] Montgomery D. C.: *Design and analysis of experiments*, John Wiley & Sons, New York 2006
[36] Myers R. H., Montgomery D. C.: *Response surface methodology: process and product optimization using designed experiments*, John Wiley & Sons, New York 1995
[37] Niederliński A.:*Systemy komputerowe automatyki przemysłowej. Tom 2, Zastosowania*, WNT, Warszawa 1985
[38] Niedoba J., Niedoba W.: *Równania różniczkowe cząstkowe i zwyczajne. Zadania z matematyki*, Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, Kraków 2001
[39] Pilkey D. W., Pilkey F. D.: *Peterson's Stress Concentration Factors*, John Wiley & Sons, New Jersey 2008
[40] Povstenko J.: *Wprowadzenie do metod numerycznych*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2002
[41] Rafajłowicz E.: *Optymalizacja eksperymentu z zastosowaniami w monitorowaniu produkcji*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2005
[42] Pozrikidis C.: *Introduction to Theoretical and Computational Fluid Dynamics*, Oxford University Press, New York 1997
[43] Rakowski G.: *Metoda elementów skończonych wybrane problemy*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 1996
[44] Ralston A.: *Wstep do analizy numerycznej*, PWN, Warszawa, 1983
[45] Strzałkowski A., Śliżyński A.: *Matematyczne metody opracowywania wyników pomiarów*, PWN, Warszawa 1978
[46] Środka W.: *Trzy lekcje metody elementów skończonych*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2004

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

[47] Tsitouras Ch.: *Runge–Kutta pairs of order 5(4) satisfying only the first column simplifying assumption*, Computers & Mathematics with Applications, vol. 62, 2011, s. 770–775, doi=10.1016/j.camwa.2011.06.002
[48] Zboś D.: *Metody numeryczne: skrypt dla studentów wyższych szkół technicznych do przedmiotu: Podstawy ETO i Informatyki*, Politechnika Krakowska, Kraków 1992
[49] Zieliński R.: *Wybrane zagadnienia optymalizacji statystycznej*, PWN, Warszawa 1982
[50] Zohdi T.I., Wriggers P.: *An Introduction to Computational Micromechanics*, Lecture Notes in Applied and Computational Mechanics, Springer, Berlin Heidelberg 2008

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Skorowidz

A
A-stabilność, 116
abacus, 16
Adamsa–Bashfortha metoda, 141
Adamsa–Moultona metoda, 141
Aitkena metoda, 46
analiza
    czynnikowa, 303
    regresji, 303
aproksymacja
    integralna, 171
    przedziałowa, 171
    punktowa, 171

B
baza interpolacji, 25
błąd
    człowieka, 21
    danych wejściowych, 18
    globalny, 116
    lokalny, 116
    maszyny, 22
    metody, 19
    metody Simpsona, 77
    miary, 22
    modelu, 22
    obcięcia, 19
    zaokrągleń, 20
Boxa–Behenkena plan, 345
Boxa–Wilsona
    metoda, 320
    plan, 339
Butchera tablice, 137

C
całka Duhamela, 255
całkowanie
    ekstrapolacja Richardsona, 71
    metoda
        prostokątów, 68
        Romberga, 73
        Simpsona, 75
        trapezów, 70
    symboliczne
        Maxima, 272
        SageMath, 284
Choleskiego
    rozkład, 160
    wstępne przekształcenie, 168
Crouta rozkład, 160
czynniki, 301

D
DAKOTA, 347
Doolittle'a rozkład, 160
doświadczenie, 294
    centralne, 335, 339
    gwiezdne, 335
Duhamela całka, 255
dywergencja, 195

E
eksperyment, 303
    bierny, 303
    czynny, 303
    planowanie, 304
ekstrapolacja Richardsona, 71
element
    dostosowany, 212
    niedostosowany, 212
    paraboliczny, 241
    sześcienny, 241

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

zgodny, 212
zupełny, 212
estymacja
nieparametryczna, 298
parametryczna, 298
przedziałowa, 298
punktowa, 298
estymator
nieobciążony, 298
obciążony, 298
Eulera
metoda, 120
jawna, 120
niejawna, 120
ulepszona, 129

F
fundamentalne rozwiązanie, 193, 197
funkcja
błędu, 299
celu, 302
kształtu, 211
regresji, 311
funkcjonał, 215

G
Galerkina metoda, 219
Gaussa
kwadratura, 82
metoda eliminacji, 156
mnożniki, 314
Gaussa–Jordana metoda, 159
Gaussa–Kronroda kwadratura, 86
Gaussa–Legendre'a kwadratura, 84
Gaussa–Seidla metoda, 162
gradientów sprzężonych metoda, 166
Grama–Schmidta ortogonalizacja, 161

H
Heron, 52
Heuna metoda, 129
hipoteza
alternatywna, 297
prosta, 297
statystyczna, 297
złożona, 297
Hornera schemat, 16
Householdera przekształcenie, 175

I
identyfikacja, 304
iloraz różnicowy, 30
interpolacja
Lagrange'a, 34, 104
Newtona, 29
wielomianami Czebyszewa, 27
wielomianowa, 26
z węzłami Czebyszewa, 27

J
Jacobiego przekształcenie wstępne macierzy, 168

K
kodowanie
planu eksperymentu, 322
Yatesa, 322
kontrast określający, 333
korelacji współczynnik, 297
kostki Napiera, 16
kowariancja, 297
kowariancyjna macierz, 314
kształtu funkcja, 211
kwadratura, 80
Gaussa, 82
Gaussa–Kronroda, 86
Gaussa–Legendre'a, 84
Newtona–Cotesa, 80

L
Lagrange'a
interpolacja, 104
wielomiany, 34
Laguerre'a metoda, 65
lebendige Rechenbank, 16

M
m-ortonormalność, 252
macierz
bezwładności elementu prętowego, 230

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

informacyjna, 305
kosinusów kierunkowych, 237, 267
kowariancyjna, 314
obrotu, 236
sztywności, 226, 266
Vandermonde'a, 27, 173
maszyna
analityczna, 17
Harvard Mark 1, 17
Leibniza, 16
Pascala, 16
różnicowa, 16
Z3, 17
Maxima, 270
metoda
Δ^2 Aitkena, 46
A-stabilna, 141
Adamsa–Bashfortha, 141
Adamsa–Moultona, 141
babilońska, 16
bilansu losowego, 308
Boxa–Wilsona, 320
elementów
brzegowych, 34, 193
skończonych, 34, 211
eliminacji Gaussa, 156
Eulera, 120
jawna, 120
niejawna, 120
ulepszona, 129
Eulera–Cauchy'ego, 128
Galerkina, 219
Gaussa-Jordana, 159
Gaussa–Seidla, 162
Geara, 144
gradientów sprzężonych, 166
Heuna, 129
iteracji prostej, 37
jednokrokowa, 120
kwadratów łacińskich, 346
Laguerre'a, 65
Lanczosa, 111
linearyzacji, 48
Monte Carlo, 88, 346
nadrelaksacji, 43
Newmarka, 144
Newtona, 48
niejawna, 120, 141
podstawiania
w przód, 159
wstecz, 158
predyktor-korektor, 142
prostokątów, 68
punktu środkowego, 129
Rayleigha–Ritza, 215
regula falsi, 54
Romberga, 73, 273
różnic skończonych, 183
różnicowa, 117
różniczkowania wstecznego, 144
Rungego–Kutty, 128, 134
siecznych, 52
Simpsona, 75
stycznych, 48
superpozycji modalnej, 251
trapezów, 70, 141
Trefftza, 194
wielokrokowa, 140, 141
metody numeryczne, 17
mnożniki Gaussa, 314
modalna superpozycja, 251
model
fizyczny, 112
matematyczny, 112, 302
Monte Carlo metoda, 88, 346

N

nadrelaksacja, 43
Napiera kostki, 16
Newmarka metoda, 144
Newtona metoda, 48
Newtona–Cotesa kwadratura, 80
nierówność trójkąta, 152
norma
druga, 153
euklidesowa, 152, 153
Frobeniusa, 153
maksimum, 153
nieskończoność, 153
pierwsza, 153

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

norma spektralna, 153

O
odchylenie standardowe, 297
odpowiedzi obiektu, 301
ortogonalizacja Grama–Schmidta, 161

P
pakiet
  DAKOTA, 347
  Maxima, 270
  SageMath, 283
pascalina, 16
Pearsona korelacji współczynnik, 297
pierwiastki wielomianów, 64
Placketta–Burmana plan, 337
plan
  A-optymalny, 305
  Boxa–Behenkena, 345
  Boxa–Wilsona, 339
  D-optymalny, 305
  dwupoziomowy
    całkowity, 326
    ćwiartkowy, 328
    ósemkowy, 328
    połówkowy, 328
    ułamkowy, 326
  E-optymalny, 305
  G-optymalny, 305
  kompozycyjny, 335
  ortogonalny, 305
  pięciopoziomowy, 335
  Placketta–Burmana, 337
  rotatabilny, 304
  trójpoziomowy, 334
    całkowity, 334
  ułamkowy, 328
planowanie eksperymentu, 304
podstawianie
  w przód, 159
  wstecz, 158
powierzchnia odpowiedzi, 321, 323
poziom
  istotności, 298
  ufności, 298
precyzja arytmetyki, 20
predyktor-korektor metoda, 142
problem
  brzegowy, 115
  początkowo-brzegowy, 115
  początkowy, 115
  własny, 115
promień gwiezdny, 338
przedział ufności, 298
przekształcenie
  Householdera, 175
  wstępne
    Choleskiego, 168
    Jacobiego, 168
  wstępne macierzy, 168

R
rachunek wariacyjny, 215
Rayleigha tłumienie, 250, 255
Rayleigha–Ritza metoda, 215
regresji funkcja, 311
regula falsi, 54
residuum, 56, 154, 166
Richardsona ekstrapolacja, 71
Romberga metoda, 73
rozbieżność wektora, 195
rozkład
  Choleskiego, 160
  Crouta, 160
  Doolittle'a, 160
  QR, 161
  SVD, 175
  względem wartości szczególnych, 175
równania różniczkowe
  cząstkowe, 114
  eliptyczne, 114
  hiperboliczne, 115
  paraboliczne, 115
  sztywne, 144
  zwyczajne, 114
równość charakterystyczna, 333
różnic skończonych metoda, 183
różnicowy schemat, 185
różniczkowanie, 100

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356

Rungego
zasada, 148
zjawisko, 80
Rungego–Kutty metoda, 128, 134
rząd metody, 116

S
SageMath, 283
schemat różnicowy, 185
siecznych metoda, 52
Simpsona
metoda, 75
wzór złożony, 77
stabilność, 115
standaryzacja zmiennych, 323
superpozycja modalna, 251
suwak logarytmiczny, 16
SVD rozkład, 175
sztywności macierz, 226

T
tablice
Butchera, 137
iloczynowe, 15
ilorazów różnicowych, 30
odwrotności, 15
tłumienie Rayleigha, 250, 255
trapezów metoda, 70, 141
trójkąta nierówność, 152

U
ufności
poziom, 298
przedział, 298
uwarunkowanie
problemu, 18
układu równań, 154
wskaźnik, 154

W
warunki zakończenia obliczeń, 56
wielomiany
Czebyszewa, 27
Lagrange'a, 34
wskaźnik uwarunkowania, 154
wstępne
przekształcenie
Choleskiego, 168
Jacobiego, 168
wykładnik zbieżności, 50

Z
zależność stochastyczna, 310
zasada Rungego, 148
zbieżność, 116
kwadratowa, 51
liniowa, 51
logarytmiczna, 51
rzędu p, 51
sześcienna, 51
zdolność rozdzielcza planu, 328
zgodność, 118
zjawisko
kwantyzacji, 91
Rungego, 80
zmienne objaśniające, 301

Plik zabezpieczony watermarkiem jawnym i niejawnym: 10941133A3130356